Homeostasis in Neural Networks: Implications for Information Processing and Learning

DISSERTATION
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Physik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Fabian Schubert
aus Frankfurt am Main

Frankfurt am Main, Januar 2022
(D30)

Vom Fachbereich Physik der
Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Harald Appelshäuser
Gutachter: Prof. Dr. Claudius Gros
Prof. Dr. Jochen Triesch

Datum der Disputation:

Zusammenfassung

Homöostase bezeichnet die Fähigkeit eines dynamischen Systems, mittels regelnder Prozesse einen intrinsischen Zustand gegenüber äußeren Einflüssen zu stabilisieren. Derartige Prozesse finden sich insbesondere in Organismen, deren Funktion nur unter gewissen physikalischen Rahmenbedingungen gewährleistet ist. Ähnlich wie beispielsweise die Regulation der Kerntemperatur von Säugetieren finden sich auch im Gehirn Mechanismen, die die mittlere neuronale Aktivität auf einem bestimmten Niveau halten, um eine normale kognitive Funktionsfähigkeit zu ermöglichen. Als dynamische Prozesse wurden verschiedene Formen dieser neuronalen Homöostase bisher auch theoretisch mittels numerischer Modelle untersucht. Neben den biochemischen Grundlagen dieser Adaptionsprozesse stellt sich aus solch einer theoretischen Sicht auch die Frage, in welche Wechselwirkung diese im Gehirn mit weiteren dynamischen Komponenten treten.

Zum einen ist dabei von Interesse, wie sich neuronale Systeme so selbst regulieren, dass ihre dynamischen Eigenschaften komplexe Informationsverarbeitung ermöglichen. Insbesondere im Fall von zeitlich variablen sensorischen Stimuli haben die dynamischen Eigenschaften des sie verarbeitenenden neuronalen Systems einen großen Einfluss auf die Effektivität bzw. Präzision solcher Verarbeitungsprozesse. Zum anderen ist zu berücksichtigen, dass homöostatische Mechanismen Einfluss auf synaptische Plastizität und Lernprozesse nehmen können, da diese unmittelbar durch die auftretenden neuronalen Aktivitätsmuster beeinflusst werden. Eine genauere Betrachtung dieser beiden Aspekte anhand zweier theoretischer Netzwerkmodelle ist Kernthema der vorgelegten Arbeit.

Ein besonderer Aspekt von homöostatischen Adaptionsmechanismen ist die Tatsache, dass mehrere Parameter zeitgleich dynamischen Veränderungen unterliegen können. Im einfachsten Fall, der über einen einzigen Parameter hinausgeht, werden zwei Parameter als Funktion intrinsischer Zustandsgrößen verändert. Wie in vorangegangenen theoretischen Arbeiten gezeigt wurde, erlaubt eine solche duale Homöostase nicht nur die Regelung des zeitlichen Mittels bestimmer Variablen, sondern auch ihrer zeitlichen Varianz. Dadurch kann entsprechend detaillierter auf die statistische Verteilung solcher intrinsischer Variablen Einfluss genommen werden. In dieser Arbeit werden zwei verschiedene Netzwerk- bzw. Neuronenmodelle vorgestellt, und deren Zusammenspiel mit dualen homöostatischen Adaptionsmechanismen untersucht. Diese beiden Modelle stehen stellvertretend für die zwei im vorigen Absatz genannten Fragestellungen. Beim ersten Modell handelt es sich um ein rekurrentes Netzwerk, also ein neuronales System, in dem synaptische Verbindungen innerhalb

einer Population von Neuronen existieren. Dieses steht den sogenannten Feedforward-Netzen gegenüber, in denen Verbindungen nur in eine Richtung zwischen mehreren Populationen existieren. Rekurrente Netzwerke weisen aufgrund ihrer Struktur grundsätzlich eine Eigendynamik auf, die es dem System erlaubt zeitabhängige externe Signale zu verarbeiten. Anhand eines solchen rekurrenten Netzwerks untersuchen wir die Wechelwirkung der dualen homöostatischen Adaption und der Eigendynamik des Netzwerks, insbesondere im Hinblick auf die Leistungsfähigkeit des Systems bei der Speicherung und Verarbeitung externer Signale.

Das zweite Modell betrachtet exemplarisch anhand einzelner Neuronen den möglichen Vorteil dualer Homöostase im Falle der bereits erwähnten Feedforward-Netze. Aus der Tatsache, dass in Feedforward-Netzen neuronale Populationen seriell und unidirektional miteinander verknüpft sind, ergibt sich biologisch ein Problem: Bei Lernprozessen, die darauf abzielen, aus einem bestimmten Input durch solch ein serielles Netzwerk einen bestimmten Output zu generieren, müssen plastische Veränderungen global über das gesamte Netzwerk koordiniert werden. Synaptische Plastizität ist aber biologisch betrachtet ein Prozess, der an lokale physikalische Zustände gebunden ist, nämlich in erster Linie an die prä- und postsynaptische neuronale Aktivität. Die Frage ist also, wie globale Lernprozesse über lokale Plastizitätsmechanismen koordiniert werden können. Theoretische Ansätze zur Lösung dieses Problems basieren im Wesentlichen auf der Annahme, dass über weitere synaptische Verbindungen ein umgekehrter (d.h. entgegen der primären seriellen Informationsverarbeitung) Informationsfluss, bzw. Feedback-Signale existieren. In dem von uns betrachteten Neuronenmodell können wir zeigen, dass eine Kombination aus dualer Homöostase und synaptischer Plastizität es erlaubt, über ein solches Feedback-Signal die lokalen Lernprozesse zu koordinieren. Die Effektivität dieses Prozesses wird zusätzlich anhand eines Lernszenarios demonstriert, in dem das Neuron eine lineare binäre Klassifikation durchführt.

Inhalt der Kapitel

Zunächst werden in Kapitel 2 die theoretischen Grundlagen zur Modellierung einzelner Neuronen bzw. neuronaler Netzwerke erläutert. Insbesondere erfolgt hier eine Einordnung rekurrenter Netzwerkmodelle in die Theorie nicht-autonomer dynamischer Systeme. Für große neuronale Netze werden Mean-Field-Methoden zur statistischen Beschreibung der Dynamik eingeführt. Abschließend wird auf einige Grundlagen der Theorie der bereits erwähnten dualen Homöostase eingegangen und erläutert, wie ein solches duales Kontrollsystem sowohl den Erwartungswert als auch die Varianz einer zeitabhängigen Größe determiniert.

Beim ersten untersuchten Modell, welches in Kapitel 3 vorgestellt wird, handelt es sich ein sogennantes Echo-State Netzwerk, das unter die Kategorie der bereits erwähnten rekurrenten Netzwerke fällt. Echo-State Netzwerke sind in der Lage, komplexe nichtlineare Berechnungen an sequentiellem externen Input durchzuführen. Die Besonderheit von Echo-State Netzwerken liegt in der Tatsache, dass das Erlernen einer bestimmten sequenziellen Informationsverarbeitung nicht durch die

Adaption der rekurrenten synaptischen Gewichte erfolgt. Stattdessen geschieht dies nur durch die Anpassung von synaptischen Verbindungen, die auf ein oder mehrere Output-Neuronen projizieren. Die Kapazität bzw. Leistungsfähigkeit solcher Netzwerke ist dennoch stark abhängig von der grundsätzlichen Skalierung der rekurrenten synaptischen Gewichte, welche sich durch den Spektralradius der rekurrenten Gewichtsmatrix parametrisieren lässt. Da der Spektralradius aber eine globale und damit relativ aufwendig zu bestimmende Größe ist, kann ein biologisches Netzwerk diesen nicht direkt regulieren. Stattdessen wird ein heuristischer Ansatz benötigt, der auf der Ebene einzelner Neuronen die Skalierung synaptischer Gewichte anpasst. Wir führen einen lokalen, biologisch plausiblen, dualen homöostatischen Regulationsmechanismus ein („Flow Control“), der in der Lage ist, diese Skalierung der synaptischen Gewichte dynamisch zu regeln. Essentiell für Flow Control ist hier, dass der lokale rekurrente und der externe Input als separate Größen betrachtet werden. Zunächst wird die Effektivität des Adaptionsmechanismus für unterschiedliche Input-Sequenzen, die sich hinsichtlich ihrer statistischen Verteilung und Korrelation unterscheiden, betrachtet. Es zeigt sich, dass sich durch Korrelationen zwischen dem externen Input einzelner Neuronen der tatsächlich resultierende Spektralradius gegenüber dem einzuregelnden Zielwert erhöht. Eine statistische Beschreibung der rekurrenten Netzwerkstruktur erlaubt uns, einen analytischen Zusammenhang zwischen der durchschnittlichen Kreuzkorrelation der neuronalen Aktivitäten und der zu erwarteten Abweichung vom Zielwert des Spektralradius herzustellen, der sich gut mit den numerischen Ergebnissen deckt.

Zusätzlich zu diesen Ergebnissen wird auch der Effekt des Adaptionsmechanismus auf die Netzwerk-Performance untersucht. Der hierzu designte Task, an dem die Output-Gewichte trainiert werden, besteht aus einem zufällig generierten, zeitlich diskreten, binären Input-Signal, an dem aus aufeinander folgenden Werten zeitlich verzögert eine XOR-Operation durchgeführt werden soll. Diese Berechnung erfordert somit einerseits die Werte der Input-Sequenz über den Zeitraum der Verzögerung im Netzwerk zu speichern, zum anderen aber auch die Fähigkeit, eine nichtlineare Operation wie in diesem Fall XOR an diesen gespeicherten Werten durchzuführen. Die Performance des Netzwerks bestimmt sich über die Korrelation zwischen den vom Task definierten Ziel-Werten und dem tatsächlich vom Netzwerk generierten Output. Unter Variation des Zielwertes des Spektralradius als auch der Varianz des verwendeten Input-Signals ergibt sich für die Performance des Netzwerks keine starke Abhängigkeit von der Input-Varianz. Insbesondere wird die maximale Performance für einen Zielwert des Spektralradius erreicht, der unabhängig von der Varianz des Input-Signals ist.

Das zweite eingangs erwähnte Modell wird in Kapitel 5 behandelt. Motivation ist hier das überwachte Lernen in hierarchischen Netzwerken. In der theoretischen Neurowissenschaft wird nach möglichen biologischen Mechanismen gesucht, die überwachtes Lernen auf Basis gegebener physiologischer Rahmenbedingungen ermöglichen. Eine Überblick über den Stand der Forschung wird zunächst in Kapitel 4 gegeben. Die Frage ist hier, wie plastische Lernprozesse innerhalb eines gestaffelten, hierarchischen Netzwerks über lokale, biologisch plausible Mechanismen koordiniert

werden. Ähnlich wie bei Algorithmen, die als überwachte Lernverfahren für künstliche neuronale Netze entwickelt wurden (beispielsweise Backpropagation), ist aus theoretischer Sicht ein Informationsfluss vonnöten, der dem der eigentlichen Richtung der Informationsverarbeitung im Netzwerk entgegensteht. Dieser Ansatz führt in Bezug auf biologisches Lernen zu zwei Teilaspekten, die zu betrachten sind.

Erstens ist hier von Bedeutung, welche Art von Information konkret über synaptische Feedback-Verbindungen übermittelt wird und wie diese kodiert ist. Im Fall von Backpropagation wird für jedes Neuron im Netzwerk ein exakter Gradient des im Output-Layer vorhandenen Fehlers berechnet. Dies ist aus mehreren Gründen aus biologischer Sicht unplausibel. Erstens würde dies eine exakte Abstimmung bzw. Parallelität zwischen Feedforward- und Feedback-Gewichten erfordern. Zweitens können die berechneten Gradienten prinzipiell sowohl positive als auch negative Werte annehmen, was eine Kodierung über neuronale Aktivität erschwert, da diese zumindest im Sinne von Feuerraten notwendigerweise strikt positiv ist. Als mögliche Lösungen für diese beiden Punkte werden in Kapitel 4 zum einen randomisierte Feedback-Gewichte diskutiert (*random feedback alignment*), zum anderen die Möglichkeit, individuelle Zielwerte für die Aktivität einzelner Neuronen anstatt der genannten Fehler-Gradienten als Feedback zu nutzen (*target propagation*).

Ein zweiter Aspekt, der aus biologischer Sicht in Bezug auf Lernprozesse in hierarchischen Netzwerken zu betrachten ist, bezieht sich darauf, wie einzelne Neuronen sowohl Feedforward- als auch Feedback-Signale intrinsisch koordinieren und wie dadurch lokal synaptische Plastizität beeinflusst wird. Ein Ansatz, dem experimentelle Befunde hinsichtlich der mikroskopischen Anatomie von Neuronen im Kortex vorausgingen, ergibt sich aus der Tatsache, dass Pyramidenzellen oft eine sehr spezifische dendritische Morphologie aufweisen: Während ein Teil der synaptischen Verbindungen sich an Dendriten befindet, die relativ dicht am Zellkern liegen (*basal*), existiert zudem eine baumartige dendritische Struktur, die sich vertikal in höhere kortikale Schichten erstreckt (*apikal*). Messungen haben gezeigt, dass dieses entfernte dendritische Kompartiment in Teilen als eigenständige Einheit verstanden werden kann, da es auch dort möglich ist, Aktionspotentiale zu initiieren. Eine Hypothese ist, dass Feedback-Signale, die in diesem oberen dendritischen Bereich zusammenlaufen, durch diese nichtlineare Dynamik Einfluss auf die Plastizität der basalen synaptischen Verbindungen nehmen können.

In dem von uns untersuchten und in Kapitel 5 diskutierten Modell werden diese spezifischen Eigenschaften mittels zwei separater externer Inputs abgebildet, die jeweils die Stimulation in den genannten basalen und apikalen dendritischen Bereichen subsumieren. Essenziell ist dabei, dass zwei Modi neuronaler Aktivität auftreten können: Für den Fall, dass nur basaler Input vorhanden ist, kann das Neuron ein bestimmtes maximales Aktivitätsniveau erreichen, welches jedoch deutlich unter der Aktivität liegt, die sich maximal aus gleichzeitig präsentem basalen und apikalen Input ergibt.

In unseren Untersuchungen wenden wir auf die basalen Synapsen zwei biologisch motivierte Plastizitätsmechanismen an: erstens die Hebbsche Plastizität, deren Grundprinzip von Donald O. Hebb bereits Ende der 1940er Jahre postuliert wurde.

Als zweiten, alternativen Mechanismus verwenden wir die sogenannte BCM-Regel, die von Elie Bienenstock, Leon Cooper und Paul Munro im Jahr 1981 als Modell für synaptische Plastizität im visuellen Kortex entwickelt wurde. Wir können zeigen, dass sowohl Hebbsche Plastizität als auch die BCM-Regel in den basalen Synapsen dazu führt, dass das Feedback-Signal im oberen dendritischen Kompartiment als Zielsignal fungiert, welches letztlich im basalen Kompartiment reproduziert wird. Ähnlich zu Flow Control sind auch hier homöostatische Prozesse ein essentieller Bestandteil: Durch eine separate Regelung beider Input-Signale wird sichergestellt, dass das Neuron sich in dem gewünschten Arbeitsbereich befindet. Der beschriebene Lernprozess kann nur teilweise reproduziert werden, wenn anstatt des komplexeren Neuronenmodells ein einfaches punktartiges Modell verwendet wird, in dem beide Input-Ströme addiert werden. In erster Linie zeigt sich dieser Unterschied darin, dass das simplere Modell hinsichtlich des Lernprozesses anfälliger für Störsignale ist: Hebbsche Plastizität führt in punktartigen Neuronenmodellen dazu, dass sich die Gewichte entlang der Hauptkomponente des präsentierten Inputs ausrichten. Für den Fall, dass diese Hauptkomponente orthogonal zu der für die Reproduktion des Lernsignals optimalen Kombination von Gewichten ist, kann diese Rekonstruktion gestört oder komplett verhindert werden. Das aus zwei Kompartimenten bestehende Modell ist hingegen deutlich robuster gegenüber einer solchen distraktiven Komponente im basalen Input. Wir demonstrieren diesen Effekt anhand zweier Lernszenarios.

Im ersten Szenario konstruieren wir das Lernsignal im apikalen Kompartiment als eine zufällig gewichtete Superposition mehrerer zufällig generierter Zeitsequenzen. Die gleichen Signale werden auch für die Erzeugung des Inputs für das basale Kompartiment verwendet. Eine perfekte Rekonstruktion des Lernsignals ist im basalen Kompartiment dann erreicht, wenn die basalen Gewichte die für das Lernsignal vorab generierte Wichtung der Zeitsequenzen reproduzieren. Allerdings wird als störender Einfluss zusätzlich der basale Input orthogonal zu dieser optimalen Gewichtung skaliert, d.h. die Varianz erhöht. Das Ausmaß dieser Skalierung bestimmt also die Stärke der Störung. Im Falle des Punktmodells werden die beiden so generierten Inputs addiert. In diesem Szenario konnten wir im Vergleich zum Punktmodell für das Zwei-Kompartimente-Modell signifikant stärkere Störungen wählen ohne die korrekte Rekonstruktion des Lernsignals nach dem Lernprozess negativ zu beeinflussen.

Den gleichen Effekt konnten wir auch im zweiten Szenario feststellen, in dem der apikale Input ein Lernsignal für eine binäre lineare Klassifikation darstellt. In diesem Fall erfolgt die Störung über eine Skalierung des zu klassifizierenden Inputs parallel zur Hyperebene, die die korrekte binäre Klassifizierung definiert. Auch hier ergibt sich nach dem Lernprozess, dass die korrekte Klassifikation für das Zwei-Kompartimente-Modell weniger stark durch die Störung beeinflusst wird.

Insgesamt können wir dadurch also die Hypothese untermauern, dass die Morphologie pyramidaler Neuronen dazu beitragen kann, dass Feedback-Signale in spezifischer Weise auf synaptische Plastizität und damit auf Lernprozesse Einfluss nehmen. Aus biologischer Sicht ist auch hervorzuheben, dass das von uns verwendete Modell keine Lernregel benutzt, die explizit auf einem Fehler zwischen apikalem Lernsignal und der basalen Rekonstruktion basiert, sondern sich auf biologisch plausible bzw.

etablierte Plastizitätsmodelle beschränkt. Die Tatsache, dass es dennoch zu einer korrekten Rekonstruktion des Lernsignals kommt, unterziehen wir im Hinblick auf die BCM-Regel in Abschnitt 5.2.3 einer genaueren theoretischen Betrachtung. Unter bestimmten vereinfachenden Annahmen können wir zeigen, dass die BCM-Regel in Kombination mit dem verwendeten Zwei-Kompartimente-Modell einer gradientenbasierten Maximierung einer Zielfunktion entspricht, die einer Erhöhung der Korrelation zwischen basalem und apikalem Input entspricht.

Zum Schluss dieser Arbeit ordnen wir in Kapitel 6 beide betrachteten Modelle in einen gemeinsamen Kontext ein. Dabei sind zwei verbindende Merkmale zu erwähnen. Zum einen die Relevanz dualer homöostatischer Adaption (und die mit ihr einhergehende Kontrolle über die im System auftretenden Fluktuationen) für Informationsverarbeitung und synaptische Plastizität. Zum anderen lässt sich als Gemeinsamkeit beider Modelle die Tatsache nennen, dass eine Trennung funktional verschiedener Input-Kanäle erfolgt, die die beschriebenen Ergebnisse ermöglichen. Abschließend diskutieren wir einige weitere Forschungsfragen, die sich aus unserer Sicht für zukünftige Arbeiten aus den hier vorgestellten Ansätzen und Ergebnissen ableiten lassen.

In Bezug auf Flow Control, also der dualen Adaption von Echo-State Netzwerken, sollte für weitere Forschung die Anwendung in biologisch realistischeren Netzwerkmodellen im Vordergrund stehen. Dies beinhaltet zum einen eine strikte Aufteilung der neuronalen Population bzw. synaptischen Verbindungen in anregende und hemmende Verbindungen, zum anderen aber auch die Implementierung in spikebasierten Neuronen im Gegensatz zu dem hier verwendeten kontinuierlichen, ratenbasierten Neuronenmodell.

Im Hinblick auf das untersuchte Zwei-Kompartimente-Modell sollte der Fokus zukünftiger Untersuchungen hingegen darin liegen, das in dieser Arbeit diskutierte Framework in ein mehrschichtiges hierarchisches Netzwerk zu integrieren, um so dessen Potenzial in komplexeren Lernszenarien zu beurteilen.

Contents

CHAPTER 1

Introduction

An essential component in the central nervous system of mammals is the neocortex. In humans, it refers to the outer layers of neural tissue in the cerebrum, which is the largest region of the central nervous system [1, p. 415–419]. About 25% of the approximately 80 billion neurons in the human brain are located within the neocortex [2], which also makes up for about 40% of the total mass [1, p. 417], illustrating its relative importance for brain function.

A major portion of experimental and theoretical neuroscience is dedicated to understanding how the cognitive capabilities of humans and other mammals are linked to the physical processes inside the cortex. Experimental work has shaped the picture of an extremely complex structure that gives rise to electrical and chemical dynamic processes on spatial and time scales ranging from micrometers to centimeters as well as milliseconds to days, months and years. While this vast spatiotemporal range still poses a challenge to theories of the brain and the cortex in particular, there is some general consensus on the basic components that a model of cortical dynamics and function should comprise.

First, it is generally believed within neuroscience that the "mental state" of an animal is reflected in the dynamics of the neurons of its central nervous system. In particular, higher cognitive functioning, e.g. the processing and interpretation of sensory information and the planning and execution of movements are reflected in the electrical patterns observed in the cortex [3, p. 195–199]. Though this view has been the subject of many debates regarding the philosophy of the mind—that is, the nature of the "relation between the mind and the physical"—it is generally not much of a concern within the natural sciences. Therefore, any mechanistic model of the cortex intending to explain some cognitive function has to do so by predicting neural electrical activity and relating it to the function in question. On a microscopic level, the "electrical activity" of individual neurons is characterized by temporal patterns of so-called action potentials: Short, highly stereotyped spikes of the electric potential between the interior and exterior of a nerve cell. Naturally, the appearance of these events in a single neuron can vary in their overall amount, as well as their temporal regularity. A detailed introduction will be given in Section 2.2. For now, "neuronal activity" shall refer to the temporal frequency of action potentials, i.e. the "firing rate".

Second, the overall dynamics of cortical activity is considered an emergent result of the physical coupling between the neurons as well as external driving from sensory inputs. This coupling is predominantly implemented by the transmission of electrical signals via axons, synapses and dendrites, see Section 2.2. To draw an analogy from physics, it is the physical coupling between the elements of a many-body system that leads to its potentially very complex dynamical properties. The entirety of synaptic connections in the brain is generally referred to as the connectome.

Third, the coupling between neurons is not static, but subject to dynamical processes as well. This effect, generally known as synaptic plasticity, allows the brain to adapt to its environment, form memories and learn from experience [4, p. 615–618, 1493]. While some plasticity mechanisms can act on the same time scale as the dynamics of neuronal activity [5], most changes happen on a slower time scale than the dynamics of neuronal activity that are shaped by the synaptic coupling [3, p. 719].

Ultimately, the study of plasticity is driven by the belief that most changes within the connectome serve a purpose: to shape the dynamic properties of the brain in such a way that it can process sensory information from the exterior physical world, plan and execute movements and, thereby, provide an evolutionary advantage [6]. As more and more plasticity mechanisms were discovered in experiments, a conceptual distinction emerged: Some changes in the connectome appeared to be suitable for explaining the emergence of specific functionalities, e.g. the execution of motor patterns, the recognition of sensory patterns or the formation of memories [4, p. 1281, 1483]. One of the most prominent plasticity mechanisms falling into this category is known as Hebbian learning [4, p. 1498]: Neurons that are synchronously active form stronger synaptic connections. Other adaptations seemed to be less specific in their function, but rather served the more general purpose of maintaining certain statistical properties of neuronal activity over time. The latter, known as homeostatic plasticity, was first described in the early 1990s, where it was found that neurons that were forced into a state of higher activity autonomously returned to their previous activity level by means of compensatory processes [7, 8]. As a multitude of homeostatic controls were subsequently found experimentally, this raised further questions from a control theoretic perspective: What dynamics result from multiple—potentially conflicting—control mechanisms attempting to regulate the average neuronal activity? Theoretical studies have shown that under certain stability conditions, multiple competing control mechanisms allows for the tuning of higher moments or entire distribution of neuronal activity [9, 10, 11, 12], in contrast to the simple case of a single parameter regulating the first moment. In particular, two dynamic control parameters, referred to as dual homeostasis, can adjust the temporal mean and variance of neuronal activity [11, 12].

In this work, we will present the potential role of dual homeostasis in the context of two different frameworks: reservoir computing and supervised learning. Both of these concepts have a strong relationship to research on artificial neural networks and machine learning:

Research on reservoir computing is both concerned with applying it as a machine learning framework that can be applied to sequence processing in practical applications, as well as its possible role in understanding recurrent network dynamics and learning. Reservoir computing is approaching recurrent neural networks such as those found in the cortex as high-dimensional systems that serve as dynamical reservoirs, allowing for complex computations on external input patterns. Synaptic connections in this reservoir are not adapted to give rise to a specific dynamic behavior. Rather, homeostatic regulatory processes drive the system towards a dynamical state that is generally beneficial for signal processing. We will discuss dual dual homeostasis as a potential candidate for such a regulatory system.

Supervised learning is an ubiquitous approach in machine learning, which can be applied if the desired response of a system to a given input is known. Different theories as to what extent this learning scheme is present in the brain have been proposed [13, 14, 15, 16]. Here, we will present a model that utilizes the interplay between dual homeostasis and Hebbian learning to form a supervised learning scheme. Building upon previous research incorporating the specific compartmental anatomy and intrinsic dynamics of Pyramidal Neurons as a way of implementing a form of feedback learning, we show that the combination of Hebbian plasticity and dual homeostasis can drive a top-down supervised learning scheme.

A Note on Nomenclature

Both the term "neural" as well as "neuronal" can be found in the scientific literature. While in some cases, both terms have been used interchangeably to refer to the same concept, we make use of both terms in this work, adhering to the usual definition: While "neural" concerns nerves, "neuronal" refers to something pertaining neurons. As the models treated in our work are supposed to model properties of neurons, we will use the "neuronal" in most of the cases. An exception is the term "neural network": Albeit being inconsistent with the given definition of "neural", it is the standard term used when referring to mathematical models describing ensembles of neurons.

CHAPTER 2

Basics

2.1 Historical Context

In the nineteenth century, medical researchers began to theorize about a functional separation of brain regions, mostly based on case studies where certain brain lesions caused characteristic changes in the personality, or cognitive capabilities of patients [17]. Since then, research in humans as well as other mammals has further unveiled different functional areas across the cerebral hemisphere [3, p. 198–199]. In particular, the cortex is now typically divided into four lobes that can be associated with different cognitive functions, e.g. visual (occipital lobe), auditory (temporal lobe) and somatosensory processing (parietal lobe), as well as motor control, action planning and short term memory (frontal lobe) [18]. While these discoveries consolidated the belief that cognitive function is rooted within physical structure of the brain, they did not yet link actual physical phenomena in the brain with particular mental processes. Even though first experiments using electroencephalography (EEG) date back to the end of the nineteenth century, it was not until the 1920s that EEG measurements were used to investigate their relation to mental states [19].

On a smaller scale, the groundbreaking work by Ramón y Cajal and Golgi at the end of the nineteenth century first showed the basic microscopical structure of the human neocortex [20]: Neurons were found to be generally organized into six cortical layers, exhibiting differences in the shapes and densities of neurons [1, p. 418].

Apart from the discovery of this laminar structure, these early results also showed the highly complex anatomy of the neurons itself: Originating from the cell body, tree-like structures were observed, that could span vertically across the entire cortical tissue [21]. While these findings did suggest some form of discrete network structure between nerve cells, it was not before the advent of electron microscopy until it became completely evident that neurons are interconnected via axons and synapses that are attached to what is referred to as dendrites.

A milestone in the theoretical study of neuronal dynamics was the work by Alan Hodgkin and Andrew Huxley in the early 1950s [22]. Hodgkin and Huxley managed to map the behavior of neurons to an electrical circuit model containing a number of non-linear components. The model, consisting of a set of coupled differential equations opened up the possibility for a mathematically rigorous analysis using the tools of dynamical systems theory.

Due to the continuous refinement of experimental tools, the historical development of neuroscience progressed from macroscopic observations to increasingly finer detail. In a sense, the theoretical mathematical modeling and analysis of neural systems took the opposite route, starting at the modeling of individual neurons as done in the aforementioned works by Hodgkin and Huxley, followed by the construction of increasingly larger and complex neuronal ensembles. On a practical side, this development was also driven by the steadily increasing computing power that researchers could use for those models. In the following sections, we will take this opposite route, and lay out the theoretical basis for modeling cortical neural networks by starting at the single neuron level, followed by the synaptic coupling of multiple neurons towards statistical models of large-scale networks.

2.2 Single Neuron Models

As stated in the previous section, what makes a neuron the basic unit of cognition is its ability to exhibit complex electrical dynamics. In the following, we will describe the physical basis of these dynamics.

2.2.1 General Anatomy

Fig. 2.1 shows the basic anatomy of neurons. Although a wide anatomical variety exists, three main parts can be identified. First, a cell body, consisting of a large component, the soma, which holds the cell nucleus. Second, a large amount of branches, called dendrites. Those are connected to the axon terminals of other neurons via synapses. The number of synaptic connections per neuron in the human neocortex is roughly estimated to be on the order of 10^3 [23]. The entirety of these connections transmit action potentials from presynaptic neurons causing changes in the intracellular electric potential which can, in turn, trigger the initiation of action potentials in the soma. These are then propagated along the third major component, the axon. Shown in beige in Fig. 2.1 is the myelin sheath covering the axon. This insulating layer facilitates the propagation of electrical impulses, and its disintegration, e.g. due to diseases such as multiple sclerosis can have severely detrimental effects on motor control, sensory perception and cognition. The axon eventually separates into thinner branches which end in axon terminals, attached to the dendritic trees of other neurons. This is where electric signals are transmitted from one nerve cell to another.

2.2.2 Equilibrium Potential

Just like other cells, neurons are enclosed by an insulating membrane. This membrane prevents ions from freely floating in and out of the cell. In particular, this separation causes a difference in the concentration of ions, the most prominent being sodium (Na^+), potassium (K^+) and chloride (Cl^-) ions, see inset B in Fig. 2.1: While sodium and chloride have a higher concentration outside the cell, potassium has a higher concentration inside. This imbalance stems from a combination of

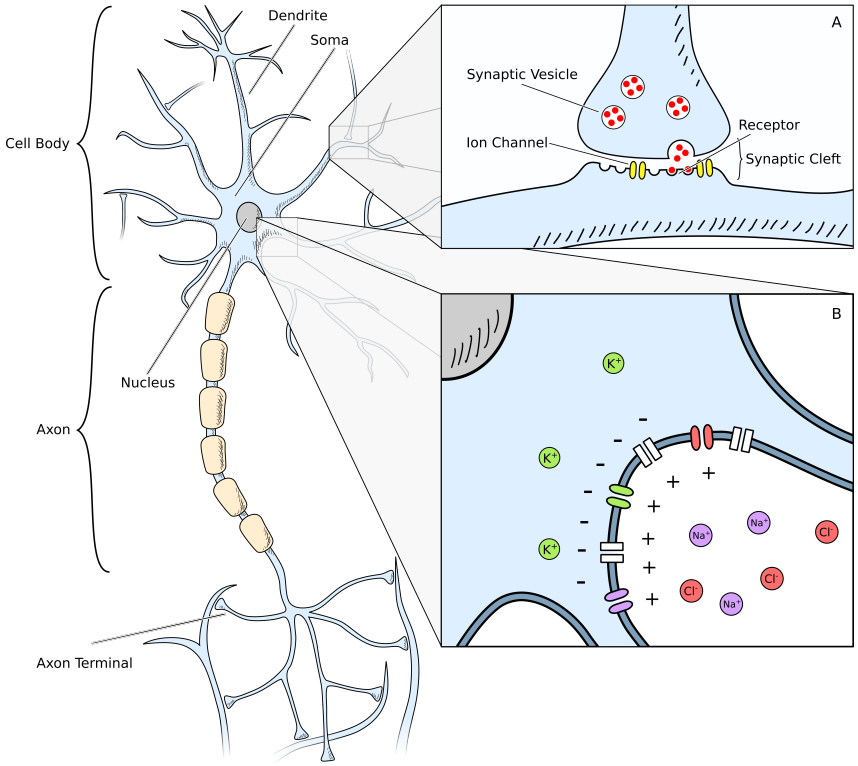

Figure 2.1: Basic anatomy of a nerve cell. The colored circles in panel B represent intra- and extracellular ions. The rectangular shapes in the cell membrane represent active ion pumps. Some ion channels are only permeable to specific ions, illustrated by the respective colored membrane insets.

effects, namely active ion transport across the membrane via so-called ion pumps (represented by rectangular shapes in inset B in Fig. 2.1), osmotic pressure and a gradient between the intracellular and extracellular potential. Ion pumps actively move ions unidirectionally from one side of the membrane to the other. For example, the sodium-potassium pump (sodium–potassium adenosine triphosphatase) transfers sodium out of the cell and potassium into the cell. As the differences in the ion concentrations across the membrane increase, the osmotic pressure increases, reducing the effectiveness of the ion transfer. Furthermore, a gradient in the electric potential arises due to the differences in concentration. Altogether, these effects generate a thermodynamic equilibrium state for a given ion type, associated with a certain difference in the intra- and extracellular electric potential. This so-called Nernst potential, if measured as $\Delta E = E_{\text{inside}} - E_{\text{outside}}$, is given by

$$\Delta E = \frac{k_{\text{B}} T}{q} \ln \frac{n_{\text{out}}}{n_{\text{in}}} \tag{2.1}$$

where k_{B} is the Boltzmann constant, T is the temperature, q the ion charge and n_{out} and n_{in} are the respective ion densities with arbitrary units, as it is their ratio that determines the resulting potential. Since this defines a state of equilibrium, it is also

called the reversal potential: A voltage difference below the reversal potential causes an inward current of positively charged ions. As an example, the reversal potential of sodium is approximately +55 mV [4, p. 153].

However, since different types of ions are present inside and outside the cell, an equilibrium is not simultaneously present for all ions. Still, one finds an intracellular potential where all ionic currents cancel to an *effective* equilibrium. This equilibrium is typically found at approximately −65 mV. As previously explained, this potential causes sodium ions to flow into the cell. On the other hand, the reversal potential of potassium is approximately −77 mV, hence potassium ions flow outward. In total, all these ionic currents result in a dynamic equilibrium.

2.2.3 Synapses

Synapses are the main component of electrical communication between neurons. While electrical synapses also exist (with a direct conducting link between neurons), the dominant type of synapse in the cortex is the chemical synapse, which transmits signals via the release of neurotransmitters. As shown in inset A in Fig. 2.1, synapses consist of two parts, the presynaptic axon terminal (top) and a region on a dendrite of the postsynaptic neuron (bottom) with specific receptors located in the membrane. When the axon terminal is depolarized, it causes a series of biochemical reactions that lead to the release of neurotransmitters from synaptic vesicles. These neurotransmitters diffuse through the synaptic cleft and bind to receptors on the postsynaptic membrane. This binding in turn causes local ion channels to open and can either cause a depolarization or hyperpolarization relative to the equilibrium potential. Various receptor types exist and their effect on the postsynaptic potential determines whether the synapse is referred to as excitatory (depolarizing) or inhibitory (hyperpolarizing). The time scale of the ionic currents caused by the synaptic transmission is generally very short at approximately 2 ms [3, p. 105]. This causes a transient change in the postsynaptic potential on the order of millivolts. It is crucial, however, that the effect on the postsynaptic potential can vary significantly among synapses, and it is this variation in synaptic efficacy that defines the structure of the connectome.

2.2.4 Intrinsic Dynamics

So far, we have described the neuron as an enclosed structure whose membrane is permeable to certain types of ions (yielding an equilibrium potential of approximately −65 mV), which can be transiently perturbed via synaptic currents. Using this picture, we can model the neuron as a capacitor in a passive electrical circuit. For this model, the total electric current I_m and the voltage across the membrane V_m are related via:

$$I_\mathrm{m} = C\dot{V}_\mathrm{m} \,. \tag{2.2}$$

Here, C is the capacity of the cell. The current I_m is the sum of the sodium and potassium currents I_Na and I_K, the synaptic currents I_S, as well as a leakage current

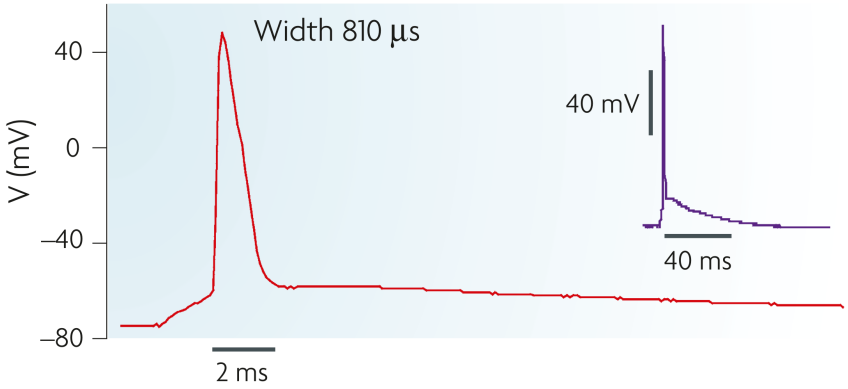

Figure 2.2: Recording of an action potential in a mouse hippocampal pyramidal neuron. Taken from Bean [24].

that subsumes other less dominant ionic currents, including chloride. Each of these currents I_i are simply given by the respective reversal potential V_i and a specific conductance g_i via

$$I_i = g_i(V_i - V_\mathrm{m}) \,, \tag{2.3}$$

which reflects the effect of the respective reversal potentials. In total, the current reads

$$C\dot{V}_\mathrm{m} = I_\mathrm{m} = g_\mathrm{K}(V_\mathrm{K} - V_\mathrm{m}) + g_\mathrm{Na}(V_\mathrm{Na} - V_\mathrm{m}) + g_\mathrm{L}(V_\mathrm{L} - V_\mathrm{m}) + g_\mathrm{S}(V_\mathrm{S} - V_\mathrm{m}) \,. \tag{2.4}$$

Note that we already described the synapses as actively changing the behavior of local ion channels over time, which can be accounted for by making the synaptic conductance g_S time dependent. Still, in its current form, the solution of (2.4) would only yield a fluctuating voltage trace, depending on the changes in the synaptic conductance. In contrast, neurons have the intrinsic ability to elicit short bursts of depolarization, known as action potentials. An example is shown in Fig. 2.2. During a very short period of approximately 2 ms, the electric potential rises to roughly 45 mV, followed by a return to the resting potential.

Hodgkin and Huxley proposed the first mathematical model explaining the emergence of action potentials [22]. It added time dependence to the sodium and potassium conductances g_Na and g_K entering (2.4). The dynamics of these conductances are described by an additional set of differential equations which, importantly, are in turn coupled to the voltage across the membrane V_m. While it provides a very accurate description of the intrinsic dynamics, it is rather complicated to analyze, which is why simplified models have been devised in an attempt to capture the essential features of the Hodgkin-Huxley model. In the following, we will discuss two such types of simplified models

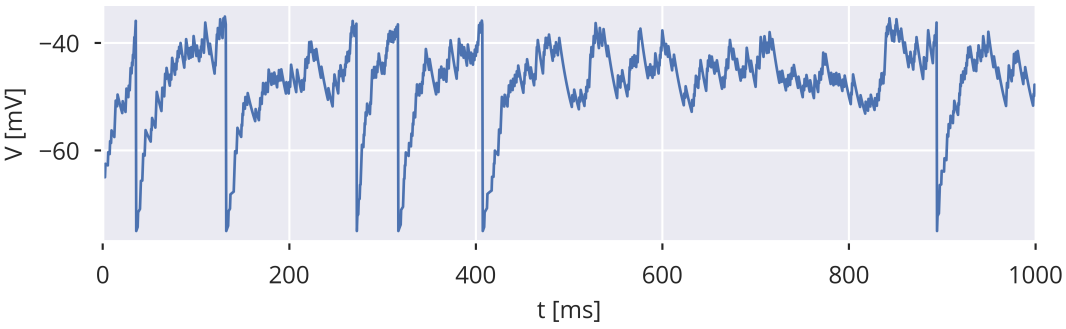

Figure 2.3: Example of the dynamics of the leaky integrate-and-fire model.

Leaky Integrate-and-Fire Model

The first observation allowing for a reduction in complexity is the fact that the temporal postsynaptic voltage response to a presynaptic action potential is very generic, apart from differences in the overall amplitude and sign accounting for the synaptic efficacy. Furthermore, these responses approximately follow a simple superposition rule, i.e. the total voltage response to all synaptic inputs is the sum of the responses to all individual presynaptic events.

Second, action potentials are typically initiated if the voltage surpasses a boundary, referred to as the firing threshold. After the spike event, the voltage returns to a certain reset potential. One of the simplest models accounting for these observations is the leaky integrate-and-fire model [25, p. 94–97]. In differential form, it can be expressed as

$$\tau \dot{V}(t) = V_{\text{rest}} - V(t) + I_{\text{syn}}(t) \tag{2.5}$$

$$I_{\text{syn}}(t) = \sum_{i,j} w_i \delta(t - t_{ij}) \, , \tag{2.6}$$

where τ is a membrane time constant, typically set to $\tau \approx 20\,\text{ms}$, and V_{rest} the resting potential. Note that, technically, the physical units of this equation are not consistent: The left hand side of the equation has the unit of volts. Thus, the term $I_{\text{syn}}(t)$ as a synaptic current is not actually an electric current, rather than an "effective" current given by $I_{\text{eff}} = I \cdot \tau / C$, with C being the membrane capacity as introduced in (2.2). This effective synaptic current $I_{\text{syn}}(t)$ is modeled as a sum over instantaneous current peaks from all synaptic connections, indexed by i and all presynaptic spiking event times t_{ij}, weighted by the synaptic efficacies w_i of the individual synapses. A spike occurs if $V(t)$ surpasses a threshold θ from below, and is directly followed by V returning to a reset voltage V_{res}. Alternatively, the return to the reset potential can be modeled by a finite-time generic voltage trace including the action potential itself. However, since only the timing of spiking events enters (2.6), the exact shape is of no importance. Still, what can be of importance is the fact that no additional spiking event can occur during this short period between the spike initiation and the return to the reset potential. An example of a voltage trace

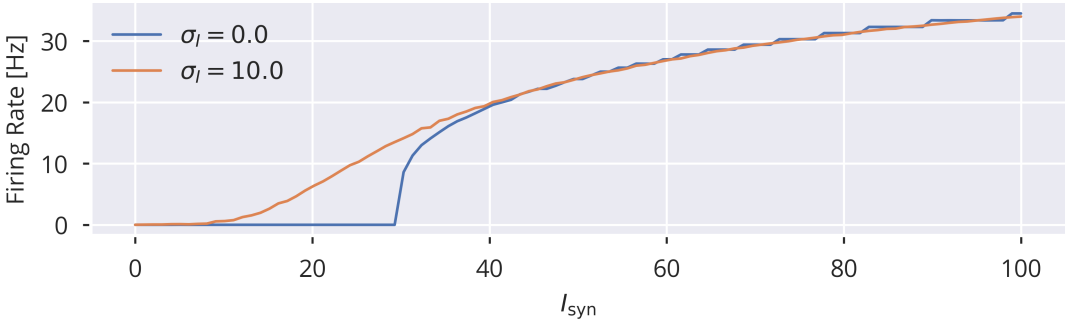

Figure 2.4: Firing rate of the leaky integrate-and-fire model as a function of the synaptic input. σ_I denotes the standard deviation of random Gaussian white noise added to the average input I_{syn}.

is shown in Fig. 2.3. In this case, the voltage is directly set back to its reset potential after passing the threshold.

Rate Models

The activity of the spiking neuron model discussed in the previous section is given by the timing of its action potentials. Using spiking models to describe the behavior of larger networks of neurons hence implicitly holds the assumption that the spiking nature of neurons is essential for an understanding of the entire system. An alternative view on the behavior of single neurons and networks is captured in what is known as rate-based models. Those are based on the observation that the frequency of spiking events in a neuron systematically varies as a function of its input current, known as the f-I curve, see Fig. 2.4. This effect has led to more abstract models, suggesting that the firing rate is the dominant unit for encoding and transmitting information in the brain [26, chapt. 15]. As one would expect, the issue of "spike versus rate" can be approached from different perspectives, potentially resulting in different conclusions in favor of or against the rate picture [27]. From an experimental standpoint, measuring firing rates has been a viable method to relate neuronal activity to perception, behavior, movement or other cognitive functions [28, 29, 30, 31]. Furthermore, large-scale measuring techniques such as functional magnetic resonance imaging or electroencephalography rely on firing rates as a neuronal physical correlate, since the observation of individual firing events is not possible with these methods. Hence, firing rates certainly *encode* information on the state of the brain, or parts of it. However, this should lead to the simple conclusion that firing rates are sufficient for explaining neuronal dynamics itself. As the theory of recurrent neural networks will be frequently discussed in this work, it is of particular interest here to examine the question of validity for rate neuron models in recurrent networks. Important work was done in this respect by Sompolinsky, van Vreeswijk and Brunel by showing that for random sparse network connectivity, a mean field theory using population firing rates is sufficient to explain the observed asynchronous irregular firing activity termed the balanced state [32, 33]. We will return to this in Section 2.3.3.

For now, we note that a typical rate model consists of a nonlinear function of the synaptic input, representing the frequency of action potentials. Obviously, since this input is generally a time dependent quantity, we face the problem of how to interpret a firing rate at any given moment in time: If we define the time averaged firing rate at time t as $\lim_{\Delta t \to 0} n[t, t+\Delta t]/\Delta t$, where $n[t, t+\Delta t]$ is the number of spikes within the time interval $[t, t+\Delta t]$, we will eventually find $n[t, t+\Delta t]$ to be zero or one, which causes the rate estimate to be either zero or diverge to infinity. There are multiple possibilities to alter our interpretation in order to avoid this issue. First, we could propose that the idea of-quasi instantaneous firing should be rejected in the first place: It is only well defined in a static situation, where the input is not changing over a prolonged time span. However, this completely rules out the possibility to use rate models in a dynamic context. As a first alternative, we could propose that the rate should be measured and averaged across an entire population of desynchronized neurons receiving the same temporal input pattern. This interpretation is reasonable in the case when we can identify populations of neurons that share similar synaptic inputs.

As a third alternative, we could propose that the instantaneous rate is the average number of spikes in a given small time window over a large number of trials: We can present the same external input pattern (if present) and average over the responses of individual neurons. This is also a typical procedure in experiments where animals or humans are given a specific task or stimulus that can be repeated multiple times [34, p. 9][35]. While this procedure can yield averaged temporal firing rate patterns of individual neurons, it comes with another set of potential issues. Simulating such a scenario with a deterministic model requires randomized initial conditions, since otherwise the resulting spike pattern would be exactly the same on each trial. This raises the question as to which extent the results are sensitive to the initial conditions: In the extreme case, slight variations in the initial conditions could result in completely different activity patterns. In a certain sense, the same problem applies to experiments: Even in the most carefully designed experiment, each trial takes place under a slightly different set of initial conditions, be it due to changes in the physical environment or the simple fact that previous trials might affect the outcome e.g. due to learning processes or exhaustion of the subject. We will return to the issue of sensitivity to initial conditions in Section 2.3.2 when discussing the dynamics of recurrent networks. Furthermore, the previously mentioned mean-field approach also utilizes a rate definition based on an average response over different initial conditions. Yet, even in the most simple interpretation of a static input, one should be aware that a perfectly constant synaptic input is not biologically plausible: Since the total synaptic current is approximately the sum of a finite amount of short individual synaptic currents, this will always result in fluctuations around a certain mean. A simple model accounting for these fluctuations is to describe these fluctuations as Gaussian white noise around a given mean. Due to the central limit theorem, this approximation is justified in the case of a large number of statistically independent presynaptic spiking events. As an example, the effect of such fluctuations on the time-averaged firing rate of a leaky integrate-and-fire model is shown in Fig. 2.4. A

finite amount of noise smoothes the f-I curve. Therefore, rate-based models usually account for this continuous transition by using continuous activation functions, most prominently the sigmoid function

$$\phi(t) = \sigma(I_{\text{syn}}(t)) = \frac{1}{1 + \exp(-gI_{\text{syn}}(t) + b)} \ . \tag{2.7}$$

This notation includes the possibility to shift the transition point using a bias b and adjust the gain, i.e. the steepness of the transition, by a factor g.

Naturally, since individual spiking events are no longer considered in this model, the synaptic input has to be calculated differently. The natural modification to (2.6) is therefore

$$I_{\text{syn}}(t) = \sum_i w_i \phi_i(t) \ , \tag{2.8}$$

where w_i still represent synaptic weights and ϕ_i are the presynaptic firing rates, including neurons acting as potential external input sources.

Technically, if none of the neurons in a network are recurrently coupled, (2.7) and (2.8) are sufficient to describe the rates within the network. However, if neurons couple recurrently, this formulation, where rates are defined as an instantaneous function of the input current, causes an issue of self-consistency. Therefore, a dynamic rate model is required. For continuous-time models, one usually chooses to model firing rates as a function of a passive membrane variable $x(t)$ with a specific time constant, whose dynamics are similar to (2.5):

$$\phi_i(t) = \sigma(x_i(t)) \tag{2.9}$$
$$\tau \dot{x}_i(t) = I_i(t) - x_i(t) \tag{2.10}$$
$$I_i(t) = \sum_j w_{ij} \phi_j(t) \ . \tag{2.11}$$

Note the additional indexing, making w_{ij} the synaptic weight between the j-th presynaptic neuron to the i-th postsynaptic neuron. Alternatively, if the network is modeled as a discrete-time dynamical system, one usually uses

$$\phi_i(t) = \sigma(x_i(t)) \tag{2.12}$$
$$x_i(t) = I_i(t) \tag{2.13}$$
$$I_i(t) = \sum_i w_{ij} \phi_j(t-1) \ . \tag{2.14}$$

Before we turn towards a more detailed description of the overall network dynamics, we note that equations (2.12) – (2.14) can be transformed into a system of

similar form but with a hyperbolic tangent as the activation function by defining:

$$\tilde{x}_i \equiv \frac{x_i(t)}{2} - \frac{w_{ij}}{4} \tag{2.15}$$

$$\tilde{w}_{ij} \equiv w_{ij}/4 \tag{2.16}$$

$$\tilde{b}_i \equiv -\frac{\sum_j w_{ij}}{4} \tag{2.17}$$

which leads to

$$\tilde{x}_i(t) = \sum_j \tilde{w}_{ij} \tanh\left(\tilde{x}_j(t-1) - \tilde{b}_j\right) . \tag{2.18}$$

Unless stated otherwise, we will use the hyperbolic tangent as our standard activation function.

2.3 Recurrent Networks

Before introducing the theoretical basics of recurrent neural network dynamics, we shall introduce the mathematical model in its most general terms. We describe recurrent networks as a set of N neurons, that are coupled by a matrix $\widehat{W} \in \mathbb{R}^{N\times N}$, with entries $w_{ij} \in \mathbb{R}$ specifying the connection strength from neuron j to neuron i. Diagonal entries are set to $w_{ii} = 0$, i.e. self-coupling is not present. Optionally, N_{ext} external neurons can be included into the model using a weight matrix $\widehat{W}_{\text{ext}} \in \mathbb{R}^{N\times N_{\text{ext}}}$. Throughout this work, we will model neural networks as time-discrete systems. Neuronal activities are represented as a time series of vectors, $\boldsymbol{y}(ta) \in \mathbb{R}^N$ and $\boldsymbol{y}_{\text{ext}}(t) \in \mathbb{R}^{N_{\text{ext}}}$. The internal activities $\boldsymbol{y}(t)$ are given by a neuronal activation function $\phi(\cdot)$ via $\boldsymbol{y}(t) = \phi\left(\boldsymbol{x}(t) - \boldsymbol{b}\right)$, where we refer to $\boldsymbol{x}(t)$ as the membrane potential and $\boldsymbol{b}$ as the bias. The activation function is applied element-wise to the vector. The membrane potential is then given by a sum of the projection of the internal and external activity. In vector notation, the dynamical system is thus given by

$$\boldsymbol{y}(t) = \phi\left(\boldsymbol{x}(t) - \boldsymbol{b}\right) \tag{2.19}$$

$$\boldsymbol{x}(t) = \widehat{W}\boldsymbol{y}(t-1) + \widehat{W}_{\text{ext}}\boldsymbol{y}_{\text{ext}}(t-1) . \tag{2.20}$$

In addition, the recurrent neurons of the network might project onto another set of neurons, which we shall interpret as the output of the network:

$$\boldsymbol{x}_{\text{out}}(t) = \widehat{W}_{\text{out}}\boldsymbol{y}(t) . \tag{2.21}$$

An illustration of the entire architecture is shown in Fig. 2.5.

2.3.1 Biological Motivation of the Network Architecture

As we shall see, the relative weight of each class of the synaptic connections has significant effects on the behavior of the system. This raises the question of the

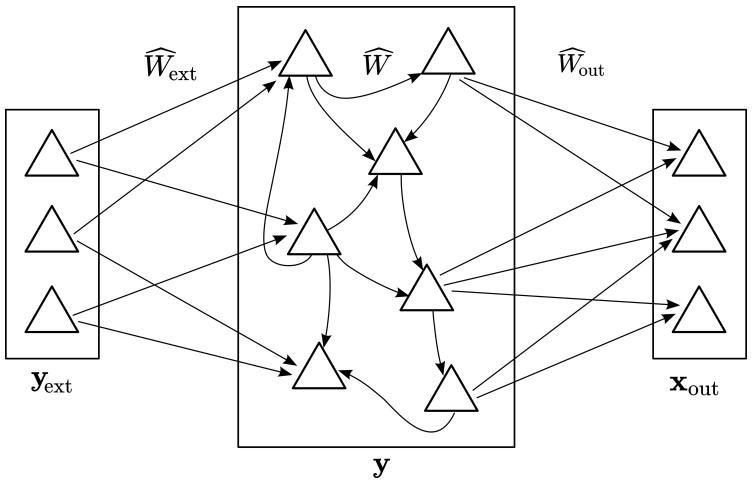

Figure 2.5: Illustration of the recurrent network structure as defined by (2.19)–(2.21).

actual importance of each of these elements in a biological context. Obviously, the recurrent network model presented here is a strong abstraction from biological cortical networks. Yet, similar models have been widely used and accepted as models for cortical networks [36, 37] and their success in explaining temporal dynamics, or cognitive functions involving time, suggests that recurrence plays a significant role in brain function. Furthermore, experimental measurements indeed indicate an abundance of local recurrent synaptic connections among various cortical regions [38].

A well-known cortical network model is referred to as the *canonical microcircuit* [39]. Speaking in broader terms, it supports the notion of the cortex as a network of strong local, intralaminar connections, as well as additional "external" input, mainly from thalamic neurons or from other more distant cortical regions. Naturally, in the generic model presented here, the former would thus be subsumed by the matrix $\widehat{W}$, while the latter corresponds to $\widehat{W}_{\text{ext}}$.

2.3.2 Recurrent Neural Networks as Dynamical Systems

Equations (2.19) and (2.20) define a particular realization of a mathematical structure known as a discrete dynamical system. The mathematical theory of dynamical systems allows us to gain insights into the behavior of the physical system we attempt to model: classifying and quantifying different dynamical states and, crucially, relate those to the parameters that enter our model.

This section will introduce the necessary theory and exemplify it using the introduced recurrent network model. As it uses discretized time steps, we will focus on the theory of discrete-time dynamical systems. The theory introduced in this section can be found in various textbooks on dynamical systems. As a popular choice, we refer to Strogatz [40] for a rigorous introduction.

Maps

Discrete-time dynamical systems are also known as maps, and we formally define them by a function

$$\boldsymbol{x} \mapsto G(\boldsymbol{x}) \,, \quad \boldsymbol{x}, G(\boldsymbol{x}) \in \mathbb{R}^n \tag{2.22}$$

that allows us to construct a flow Φ as

$$\Phi(\boldsymbol{x}_0, t) \equiv G^{(t)}(\boldsymbol{x}_0) \,, \quad \boldsymbol{x}_0 \in \mathbb{R}^n \,, t \in \mathbb{N}_0 \tag{2.23}$$

where $G^{(t)}$ represents the t-times iterated application of G. The n-dimensional Euclidean space in which the state of the system is represented shall be referred to as the phase space of the dynamical system. Technically, there is no need for G to be differentiable with respect to $\boldsymbol{x}$. However, for practical purposes, we will assume that the differential of G with respect to $\boldsymbol{x}$ exists everywhere, unless stated otherwise. As a shorthand notation, we usually refer to $\Phi(\boldsymbol{x}_0, t)$ simply by $\boldsymbol{x}(t)$, keeping the possibility in mind that there is, in general, an explicit dependence on the initial conditions.

Note that the flow introduced here defines an *autonomous* dynamical system: time only enters as an incremental parameter that represents the progression of the system through time. The iterative map G is time independent. If this is not the case, the system is non-autonomous and can be defined by a time dependent function

$$(\boldsymbol{x}, t) \mapsto G(\boldsymbol{x}, t) \,, \quad \boldsymbol{x}, G(\boldsymbol{x}, t) \in \mathbb{R}^n \,, \; t \in \mathbb{Z} \tag{2.24}$$

and the corresponding flow Φ which is now implicitly defined as

$$\Phi\left(\boldsymbol{x}_0, t_0, t\right) \equiv G\left(\Phi\left(\boldsymbol{x}_0, t_0, t-1\right), t-1\right) \,, \quad t > t_0 \tag{2.25}$$

$$\Phi\left(\boldsymbol{x}_0, t_0, t_0\right) \equiv \boldsymbol{x}_0 \,. \tag{2.26}$$

In contrast to the autonomous case, both t and t_0 now represent "absolute" time: For the autonomous case, t simply represents the number of iterations on the initial state. For the non-autonomous system, however, the explicit dependence of G on some absolute time t also requires the flow to be placed within this absolute time reference.

In the absence of external input, our recurrent network model corresponds to an autonomous system, that is

$$\boldsymbol{y}(t) = \phi\left(\widehat{W}\boldsymbol{y}(t-1) - \boldsymbol{b}\right) \,. \tag{2.27}$$

External activity $\boldsymbol{y}_{\text{ext}}$ then adds an explicit time dependence to the map.

The mathematical treatment of autonomous and non-autonomous systems differs in a number of aspects. We will first introduce the required basics of autonomous systems and then cover the relevant theory for the non-autonomous case.

Fixed Points

We define the fixed point of an autonomous discrete dynamical system as a point $\boldsymbol{x}^* \in \mathbb{R}^n$ in phase space that satisfies

$$G(\boldsymbol{x}^*) = \boldsymbol{x}^* \,, \tag{2.28}$$

that is, the map G maps $\boldsymbol{x}^*$ onto itself. By induction and (2.23), it is evident that the flow of the system will remain in $\boldsymbol{x}^*$ for all times if $\boldsymbol{x}_0 = \boldsymbol{x}^*$.

The *stability* of a fixed point $\boldsymbol{x}^*$ is defined by the long term behavior of the dynamical system in the neighborhood of $\boldsymbol{x}^*$:

A fixed point $\boldsymbol{x}^*$ is *asymptotically stable* if there exists an $\epsilon > 0$ for which any initial condition satisfying $|\boldsymbol{x}_0 - \boldsymbol{x}^*| < \epsilon$ leads to $\Phi(\boldsymbol{x}_0, t) \to \boldsymbol{x}^*$ as $t \to \infty$.
A fixed point $\boldsymbol{x}^*$ is *unstable* if for arbitrarily small $\epsilon > 0$ and all initial conditions $\boldsymbol{x}_0$ with $0 < \|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq \epsilon$, there is a t for which $\|\Phi(\boldsymbol{x}_0, t') - \boldsymbol{x}^*\| > \epsilon$ for all $t' \geq t$, where $\|\cdot\|$ denotes the Euclidean norm.

This definition formalizes the statement that fixed points are referred to as stable if small perturbations from the fixed point vanish over time, whereas for unstable fixed points, they are amplified. According to the Hartman-Grobman theorem [41], the behavior of the system close to a fixed point can be characterized by means of its first order linearization around the fixed point:

$$G(\boldsymbol{x}) = \boldsymbol{x}^* + \widehat{J}_G(\boldsymbol{x}^*)\boldsymbol{\delta} + \mathcal{O}(\boldsymbol{x}^2) \tag{2.29}$$

$$\boldsymbol{\delta} \equiv (\boldsymbol{x} - \boldsymbol{x}^*) \tag{2.30}$$

$$\left(\widehat{J}_G(\boldsymbol{x})\right)_{ij} \equiv \frac{\partial G_i(\boldsymbol{x})}{\partial x_j} \,, \tag{2.31}$$

where we have defined $\widehat{J}_G(\boldsymbol{x}^*)$ as the Jacobian matrix evaluated at $\boldsymbol{x}^*$, and $\mathcal{O}(\boldsymbol{x}^2)$ being higher order terms in $\boldsymbol{x}$. Then, the dynamics of the perturbation $\boldsymbol{\delta}(t) = \boldsymbol{x}(t) - \boldsymbol{x}^*$ is given by

$$\boldsymbol{\delta}(t+1) = \widehat{J}_G(\boldsymbol{x}^*)\boldsymbol{\delta}(t) \,. \tag{2.32}$$

The solution of this linear map is simply given by the eigenvectors $\boldsymbol{v}_i$ and eigenvalues λ_i of $\widehat{J}_G$ via

$$\boldsymbol{\delta}(t) = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i \lambda_i^t \,, \tag{2.33}$$

where the factors α_i are related to the initial condition by $\sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i = \boldsymbol{x}_0$. Writing the eigenvalue terms in their Euler representation $\lambda_i^t = r_i^t \exp(i\phi_i t)$, we see that their long term behavior is determined by their absolute values r_i. For $t \to \infty$, the term in corresponding to the eigenvalue with the largest r_i dominates in the sum of (2.33). Hence, the fixed point is stable if the largest absolute value of all eigenvalues is smaller (stable fixed point) or larger than one (unstable fixed point). This value is also referred to as the *spectral radius* of the matrix.

For two-dimensional systems, fixed points can generally be divided into five different classes, based on their eigenvalue spectrum:

- *Saddle-nodes* have two real eigenvalues, one of which is positive and one is negative.
- *Stable and unstable nodes* have real eigenvalues that are either both negative (stable) or positive (unstable).
- *Stable and unstable spirals* have complex pairs of eigenvalues whose real part is either negative (stable) or positive (unstable). In contrast to the stable and unstable nodes, the imaginary component reflects a rotational component in the dynamics.

Taking the recurrent network model with $\phi(\boldsymbol{x}) = \tanh(\boldsymbol{x})$ as an example, fixed points of the autonomous system are given by the solution of

$$\boldsymbol{y}^* = \tanh\left(\widehat{W}\boldsymbol{y}^* - \boldsymbol{b}\right) . \tag{2.34}$$

A general analytical solution to this equation does not exist, but if we choose our biases to be zero, a trivial solution is $\boldsymbol{y}^* = 0$, which yields a Jacobian that is equivalent to the synaptic weight matrix $\widehat{W}$. The stability of the network around this fixed point is therefore determined by the spectral radius of $\widehat{W}$, which will be a reoccurring quantity throughout this thesis.

While the overall stability of a fixed point is determined by the spectral radius of its Jacobian, a further characterization is possible using the eigenvalue spectrum. For the linearized system, we can separate the eigenvalues of the Jacobian into three sets, based on whether their absolute values are smaller, equal or larger than one. if we denote the corresponding sets of eigenvectors by $S_<$, $S_=$ and $S_>$, the subspaces spanned by the respective sets of vectors are called the stable, center and unstable eigenspace of the fixed point. Fig 2.6 shows an example of a three-dimensional system with a two-dimensional stable eigenspace and a one-dimensional unstable eigenspace. The dimension and orientation of these manifolds allow us to better understand the characteristic dynamics of the system close to the fixed point.

Invariant sets and Attractors

Stable fixed points as introduced in the previous section are a particular realization of the more general notion of attractors. To explain the idea behind attractors, we introduce an important concept of dynamical system theory, namely invariant sets (see, e.g. Kloeden and Rasmussen [42, p. 4]):

For a dynamical system given by the flow $\Phi(\boldsymbol{x}, t)$, the set M is called an *invariant set* of Φ if $\Phi(M, t) = M$ for all t. We denote by $\Phi(M, t)$ the set of all points generated by applying $\Phi(\cdot, t)$ to all elements in M: $\Phi(M, t) \equiv \{\Phi(m, t) : m \in M\}$.

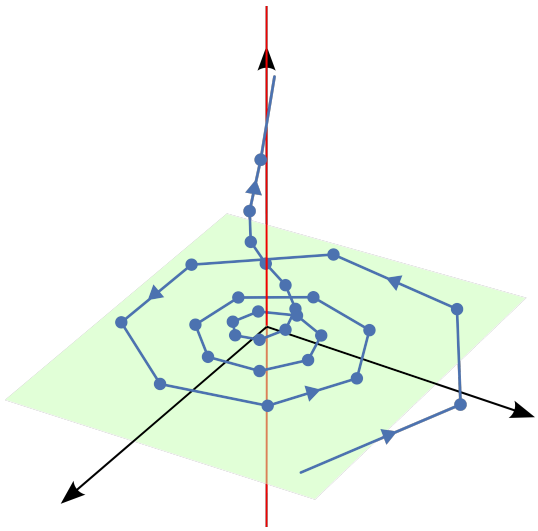

Figure 2.6: Three-dimensional discrete dynamical system with its unstable eigenspace shown in red and the stable eigenspace shown in green.

An *attractor* in the most general sense is an invariant set towards which the dynamical system evolves for some number of initial states. The geometry of these invariant sets may vary, but certain types can be identified. The simplest one, a stable fixed point, consists of a single point in phase space. Furthermore, the attractor can consist of a finite number of points (in the case of discrete systems), known as limit cycles, or infinite sets of points, with complex fractal structure, also known as strange attractors, which are indicators of chaotic behavior.

In addition to the attracting set itself, we can associate to it a so-called *basin of attraction*: It consists of all points in phase space from which the system will converge to the attractor. Therefore, the size of the basin of attraction is, in a sense, a measure for the robustness of the attractor against perturbations.

Limit cycles in discrete dynamical systems are sets of two or more points in phase space that map onto each other to form a periodic dynamical pattern. A set of unique points $\{\boldsymbol{x}_0, ..., \boldsymbol{x}_{N-1}\}$ is a limit cycle of length N of a map G if the flow satisfies $\Phi(\boldsymbol{x}_i, N) = \boldsymbol{x}_i$ and $G(\boldsymbol{x}_i) = \boldsymbol{x}_{\mod(i+1,N)}$. By defining a map $G'(\boldsymbol{x}) \equiv \Phi(\boldsymbol{x}, N)$, we find that we can analyze the stability of the limit cycle by analyzing the stability of G' at any of the points of the limit cycle. For example, the Jacobian $\widehat{J}'$ of the limit cycle evaluated at $\boldsymbol{x}_0$ is simply given by

$$\widehat{J}'(\boldsymbol{x}_0) = \widehat{J}(\boldsymbol{x}_{N-1})\,\widehat{J}(\boldsymbol{x}_{N-2})\,...\,\widehat{J}(\boldsymbol{x}_0)\ , \tag{2.35}$$

where we have used the chain rule. To see that the eigenvalues of $\widehat{J}'$ do not change across the elements of the limit cycle, we can assume that λ is an eigenvalue of $\widehat{J}'(\boldsymbol{x}_0)$

with an eigenvector $\boldsymbol{v}$ and write

$$\widehat{J}'(\boldsymbol{x}_0)\,\boldsymbol{v} = \lambda \boldsymbol{v} \tag{2.36}$$

$$\widehat{J}(\boldsymbol{x}_0)\,\widehat{J}'(\boldsymbol{x}_0)\,\boldsymbol{v} = \lambda \widehat{J}(\boldsymbol{x}_0)\,\boldsymbol{v} \tag{2.37}$$

$$\widehat{J}(\boldsymbol{x}_0)\,\widehat{J}(\boldsymbol{x}_{N-1})\,\ldots\,\widehat{J}(\boldsymbol{x}_0) = \lambda \widehat{J}(\boldsymbol{x}_0)\,\boldsymbol{v} \tag{2.38}$$

$$\widehat{J}'(\boldsymbol{x}_1)\,\boldsymbol{v}' = \lambda \boldsymbol{v}', \qquad \boldsymbol{v}' \equiv \widehat{J}(\boldsymbol{x}_0)\,\boldsymbol{v}\,. \tag{2.39}$$

Therefore, by induction, λ must be an eigenvalue of all $\widehat{J}'(\boldsymbol{x}_i)$—even though their eigenvectors will be different.

Strange attractors typically exhibit chaotic behavior (even though counterexamples exist) [43]. The fractal geometry is a consequence of the nonlinear dynamics, generating irregular sequences of states. The most important characteristic of chaotic behavior is the sensitivity of the system to its initial conditions. A similar effect was described for the local behavior around unstable fixed points, where small perturbations are amplified exponentially. However, unstable fixed points are not attractors. This emphasizes the crucial property of chaotic attractors: The attractor has a basin of attraction from which the system evolves towards the attractor, while states within the attractor diverge.

One way of analyzing the behavior of the system on a strange attractor is by means of the *Lyapunov Spectrum*. It is a natural extension of the analysis described for limit cycles, that is, in the case of an infinite series of non-repeating states generated by the system. There are different equivalent definitions of the Lyapunov spectrum. In analogy to the procedure in the case of limit cycles and (2.35), we consider an initial state $\boldsymbol{x}_0$ on an attractor and the corresponding series of states defined by the evolution of the dynamical system $\Phi(\boldsymbol{x}_0, t) = \boldsymbol{x}_t$. By denoting

$$\widehat{J}_t \equiv \widehat{J}(\boldsymbol{x}_t)\,\widehat{J}(\boldsymbol{x}_{t-1})\,\ldots\,\widehat{J}(\boldsymbol{x}_0)\ , \tag{2.40}$$

we define the Lyapunov spectrum l_i of the attractor as

$$l_i \equiv \lim_{t\to\infty} \frac{1}{2t} \ln\left(\lambda_i\left(\widehat{J}_t^\dagger \widehat{J}_t\right)\right) \tag{2.41}$$

where $\widehat{J}_t^\dagger$ is the conjugate transpose of $\widehat{J}_t$ and $\lambda_i(\widehat{J}_t^\dagger \widehat{J}_t)$ is the i-th element in the eigenvalue spectrum of $\widehat{J}_t^\dagger \widehat{J}_t$. By definition, the exponents are real valued, and the sign of the largest Lyapunov exponent $l_{\max}$ determines the long-term behavior on the attractor: A negative $l_{\max}$ implies that small perturbations decay exponentially, while $l_{\max} > 0$ is associated with an exponential growth of perturbation, which is an indicator of chaotic dynamics.

The definition of the Lyapunov spectrum generally applies to all types of attractors, including fixed points and limit cycles. In this context, it is worth to note that (2.41) contains the singular values s_i of $\widehat{J}_t$, since $\lambda_i(\widehat{J}_t^\dagger \widehat{J}_t) = s_i^2(\widehat{J}_t)$. This might appear incommensurate with the fact that we introduced the absolute values of the

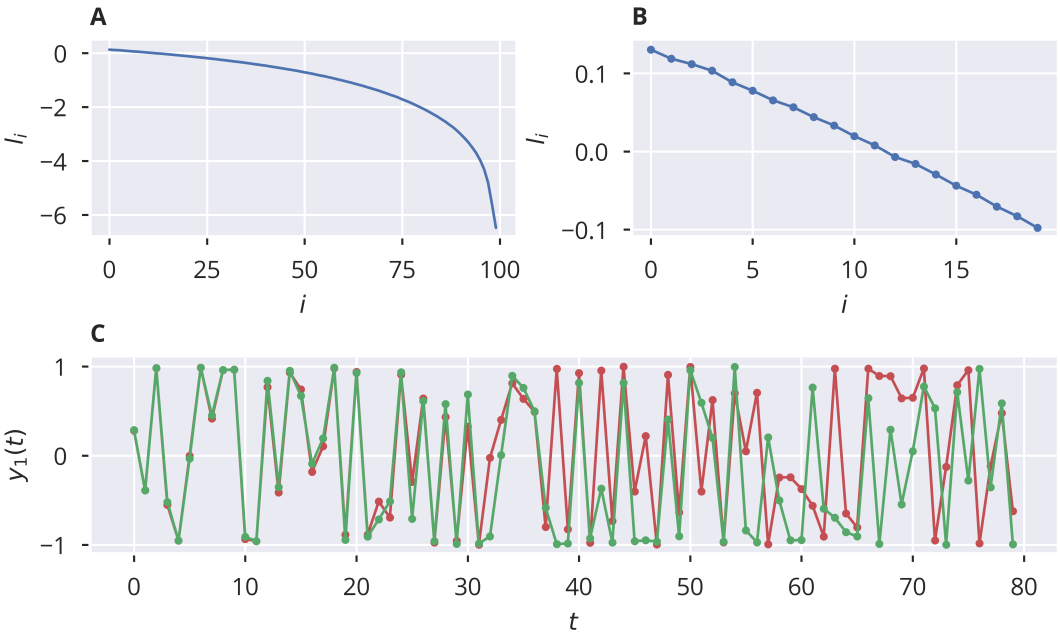

Figure 2.7: Chaotic dynamics in a recurrent network of 100 randomly connected neurons. A/B: Lyapunov spectrum of the chaotic attractor, panel B showing only the 20 largest exponents. C: Diverging trajectories of the activity of the first neuron in the network with an initial perturbation of 10^{-2}.

Jacobian as a measure for the stability of fixed points and limit cycles. The connection between both approaches lies in the following relation [44]:

$$\lim_{n\to\infty} s_i \left(\widehat{A}^n\right)^{1/n} = \left|\lambda_i \left(\widehat{A}\right)\right| . \tag{2.42}$$

For a fixed point attractor, we simply have $\widehat{J}_t = \widehat{J}^t(\boldsymbol{x}^*)$, which, in combination with (2.42), reduces (2.41) to $l_i = \ln(|\lambda_i(\widehat{J}(\boldsymbol{x}^*))|)$, showing that the largest Lyapunov exponent is just the logarithm of the spectral radius of the Jacobian of the fixed point.

Figure 2.7 illustrates chaotic behavior in a recurrent neural network of 100 randomly connected neurons. 15 of the 100 exponents are positive, causing small perturbations to be amplified.

Bifurcations

Up to this point, we treated maps as static functions that lead to certain dynamical properties. Introducing additional parameters affecting the shape of the map allows us to study the behavior of the dynamics with respect to these variables. Less substantial effects could for example be gradual changes in the shape of attractors, without fundamentally changing the dynamics. However, it is also possible to

construct parameterizations that fundamentally change the topology of the system. Such changes are referred to as *bifurcations*.

Due to the various types of bifurcations, giving a general definition to the term is not straightforward. For our purposes, a sufficiently rigorous definition can be given following Guckenheimer and Holmes [45, p. 119]:

Given a map $G(\boldsymbol{x}, \mu)$ that is parameterized by $\mu \in \mathbb{R}^d$, a value μ_0 of μ for which the flow is not structurally stable is a *bifurcation value* of μ. We informally define a system to be structurally stable with respect to the parameter μ if small changes in μ retain the topology of the system.

The number of dimensions d of μ is the *codimension* of the bifurcation, that is, the minimal number of scalar parameters necessary for the bifurcation to be observed. For most bifurcations, $d = 1$.

The most general distinction among bifurcations can be made between *local* and *global bifurcations*. Local bifurcations can be described as changes in the properties of fixed points and their local neighborhoods, whereas global bifurcations involve larger subsets of phase space, e.g. the annihilation of limit cycles. In the following, we will introduce some of the most commonly observed types. Most of the bifurcations can be found in both continuous and discrete systems, but differ in their mathematical definition. Here, we will introduce bifurcations as appearing in discrete maps. Each of these bifurcations has a *normal form*: the most mathematically simple system necessary to generate the desired behavior.

A *saddle-node bifurcation* or *fold bifurcation* is the local annihilation of a stable and an unstable fixed point. For a discrete system, its normal form can be written as $x(t+1) = x(t) + x(t)^2 + \mu$, which means that the two fixed points $x^*_{0,1} = \pm\sqrt{-\mu}$ exist for $\mu < 0$.

The *pitchfork bifurcation* shows a similar behavior, involving the annihilation of two fixed points. However, it also includes a third fixed point that changes its stability. The normal form is $x(t+1) = x(t) + \mu x(t) - x(t)^3$. The fixed point $x^*_0 =$ is always present but changes from stable to unstable as μ increases and passes $\mu_0 = 0$. Furthermore, two stable fixed points $x^*_{1,2} = \pm\sqrt{\mu}$ emerge. Similarly, we can invert the behavior by choosing the normal form to be $x(t+1) = x(t) + \mu x(t) + x(t)^3$ which gives a single unstable fixed point for $\mu > 0$, which changes to a stable fixed point for $\mu < 0$, along with the emergence of two unstable fixed points $x^*_{1,2} = \pm\sqrt{-\mu}$.

The *flip bifurcation*, or *period doubling bifurcation* only appears in discrete systems and is characterized by a period doubling of a limit cycle, where we include "period-1 limit cycles", i.e. fixed points yielding a period-2 limit cycle. A normal form of a flip bifurcation can be written as $x(t+1) = \mu\left(x^2(t) - x(t)\right)$, which has a fixed point at $x^* = 0$ that is stable for $\mu < 1$ and becomes unstable for $\mu > 1$, causing the emergence of a stable period-2 limit cycle.

The *Hopf bifurcation* is commonly defined on continuous two-dimensional dynamical systems but can also appear in discrete systems. In a Hopf bifurcation, an invariant closed curve bifurcates from a fixed point. An invariant set is defined as a subset of phase space within which a dynamical system will remain indefinitely if

initialized inside the set. One can distinguish between supercritical and subcritical Hopf bifurcations: In the supercritical case, a stable fixed point turns unstable and a stable invariant curve emerges. In the subcritical case, an unstable fixed point becomes stable and an unstable invariant curve emerges. A normal form of the discrete Hopf bifurcation can be written as

$$\boldsymbol{x}(t+1) = \left(\widehat{R}(\omega)\boldsymbol{x}(t)\right)\left[1+\alpha\left(\mu - \|\boldsymbol{x}(t)\|^2\right)\right] , \tag{2.43}$$

where $\|\cdot\|$ denotes the euclidean norm, $\widehat{R}(\omega)$ is a rotation matrix for some arbitrary angle ω and $\alpha = \pm 1$ determines whether the bifurcation, taking place at $\mu = 0$, is supercritical ($\alpha = 1$) or subcritical ($\alpha = -1$).

In the literature, global bifurcations are commonly introduced in the context of continuous systems. Still, they can also be found in discrete systems [46]. Global bifurcations emerge from the interaction of limit cycles with fixed points or other limit cycles: Similar to e.g. the pitchfork bifurcation, the dynamical properties of the system change fundamentally when those constituents come into contact in phase space.

Typical global bifurcations are for example:

- The *homoclinic bifurcation* denotes a transition point of the bifurcation parameter, where a saddle-node merges with a limit cycle and forms a homoclinic orbit: its stable and unstable manifolds connect to form a closed loop.
- The *saddle-node bifurcation of limit cycles* appears when a stable and an unstable limit cycle join and annihilate each other. This is the global equivalent to a saddle-node bifurcation.

Routes to Chaos

As mentioned in the previous section, dynamical systems such as recurrent networks can exhibit chaotic behavior for certain parameter values. As one might imagine, the onset of chaotic behavior is generally not instantaneous with respect to the bifurcation parameter. Rather, one often observes a gradual transition from regular, i.e. stationary or periodic dynamics to irregular, aperiodic behavior as parameters are varied. However, the characteristics of these transitions can vary, and different "routes to chaos" can be identified [47]. Here, we briefly discuss some of the most common types.

One of the most prominently found transition to chaos in low-dimensional systems is the *period doubling route to chaos*. Initially, a fixed point turns into a period-two limit cycle. This limit cycle loses stability again, subsequently transitioning to limit cycles with a period length twice as long as the previously stable cycle. Eventually, a value of the bifurcation parameter μ_∞ is reached where the period length diverges, corresponding to non-repeating, irregular sequences. The most prominent example of this behavior can be observed in the logistic map $x(t+1) = \mu x(t)\,(1 - x(t))$, where $\mu_\infty \approx 3.5699$. Note that the value μ_∞ is a lower bound for the possible emergence

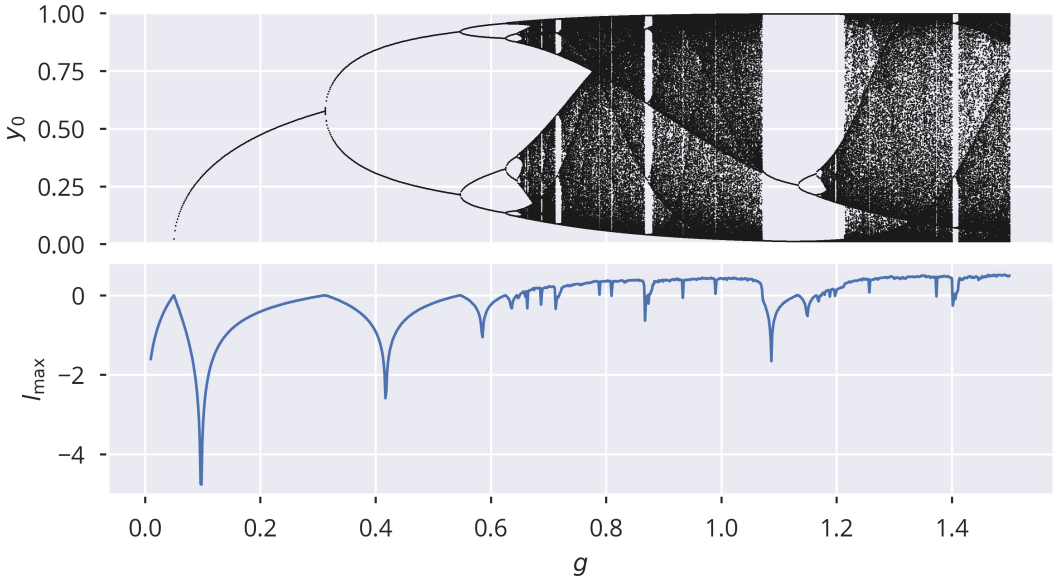

Figure 2.8: Bifurcation diagram and largest Lyapunov exponent $l_{\max}$ of a network as defined in (2.19) and (2.20), of two recurrently coupled neurons. Shown in the top panel is the projection of the activity on the first node y_0 as a function of the bifurcation parameter g, which is a scaling factor acting on the recurrent weights as $g\widehat{W}$. Values of $\widehat{W}$ are $w_{00} = -5$, $w_{01} = 5$, $w_{10} = -25$ and $w_{11} = 25$, taken from [48].

of chaos, but does not imply chaotic behavior for all $\mu > \mu_\infty$. The logistic map, for example, exhibits a return to a stable period-three cycle at $\mu \approx 3.83$ [40, p. 356]. In our recurrent network model, the same type of transition can also be observed as a function of some global scaling g on the recurrent weights, e.g. in a small network of two coupled neurons (including self coupling) [48]. Fig. 2.8 shows an example of a bifurcation diagram and the corresponding largest Lyapunov exponent. Regions of g where chaotic dynamics are present correspond to $l_{\max} > 0$, indicating sensitivity to the initial conditions.

Another type is known as the *intermittency route to chaos*, where intermittent phases of chaotic dynamics alternate with regular activity. As parameters of the system are changed, the chaotic phases become the more and more dominant, eventually leading to persistent chaotic dynamics. For example, intermittent chaos can be observed in the well-known Lorenz system [49]. Recurrent neural networks can be trained to replicate these dynamics [50].

Furthermore, the *Ruelle–Takens–Newhouse route to chaos* is characterized by two subsequent Hopf-bifurcations, yielding quasi-periodic orbits which then turn into a chaotic attractor [49]. This type of transition has been observed in large random recurrent networks [51]. In our network model, if no biases are present, the first Hopf-bifurcation appears at the $\boldsymbol{x}^* = 0$ fixed point as the spectral radius of $\widehat{W}$ becomes larger than one. We will come back to this transition when introducing the mean field theory for large random networks.

Non-autonomous Dynamical Systems

For externally driven networks, the concepts introduced for autonomous systems have to be adjusted to account for an explicit time dependence. In particular, the notion of attractors needs to be revisited.

As a first remark, we should note that a non-autonomous system defined by the map $G(\boldsymbol{x}, t)$ as introduced by (2.24)–(2.26) can always be reduced to an autonomous system by extending the phase space to $\tilde{\boldsymbol{x}} \equiv (x_1, ..., x_n, t)$, i.e. adding time as an additional coordinate. The autonomous map is then simply given by

$$\tilde{G}\left(\tilde{\boldsymbol{x}}\right) \equiv \left(G\left((\tilde{x}_1, ..., \tilde{x}_n), \tilde{x}_{n+1}\right), \tilde{x}_{n+1} + 1\right) \ . \tag{2.44}$$

However, this transformation is not particularly helpful, since it rules out the possibility of finding attractors within a finite area of phase space: The temporal coordinate will grow indefinitely.

In the following, we will extend some of the definitions given for the case of autonomous to non-autonomous systems. For a rigorous mathematical introduction, see for example Kloeden and Rasmussen [42].

Invariant non-autonomous sets are the counterpart to invariant sets introduced for autonomous systems. Instead of a single invariant set M, a series of sets

$$\mathcal{M} \equiv \{M_{t_0}, M_{t_0+1}, M_{t_0+2}, ...\} \tag{2.45}$$

is called an invariant non-autonomous set of a system $\Phi(\boldsymbol{x}, t_0, t)$ if $\Phi(M_{t_0}, t_0, t) = M_t$ for all $t \geq t_0$. Note that, as for the definition of invariant sets, we denote by $\Phi(M_{t_0}, t_0, t)$ the set resulting from applying the flow $\Phi(\cdot, t_0, t)$ to all elements in M_{t_0}.

The correspondence to invariant sets of autonomous systems can be seen by considering the special case where $M_t = M_{t'}$ for all t and t', which makes both definitions equivalent. In a sense, the definition of invariant families is less restrictive compared to invariant sets, since a family of sets fulfilling the definition can be easily obtained by choosing an initial set M_{t_0} and subsequently generating the following sets by the repeated application of the map to the last generated set. Still, the definition allows us to extend the concept of attractors to non-autonomous systems.

Generally, one can distinguish between two types of non-autonomous attractors: *Forward attractors* and *pullback attractors*.

An invariant non-autonomous set $\mathcal{M}$ of a system $\Phi(\boldsymbol{x}, t_0, t)$ is called *forward attractor* if $\lim_{t\to\infty} \mathrm{dist}\left(\Phi(\boldsymbol{x}, t_0, t), M_t\right) = 0$ for all $\boldsymbol{x}$ and t_0, where $\mathrm{dist}(\boldsymbol{x}, A)$ between a point $\boldsymbol{x}$ and a set A is defined as the smallest distance between $\boldsymbol{x}$ and all points in A.

This formal definition expresses the property of the dynamical system to converge towards the invariant set from any initial point in time and space if evolved forward.

Likewise, an invariant non-autonomous set $\mathcal{M}$ of a system $\Phi(\boldsymbol{x}, t_0, t)$ is called *pullback attractor* if $\lim_{t_0 \to -\infty} \operatorname{dist}(\Phi(\boldsymbol{x}, t_0, t), M_t) = 0$ for all $\boldsymbol{x}$ and t.

In this definition, the distance to the invariant set will tend to zero at any point in time if the starting time is pulled back to negative infinity. Based on intuition, it might appear that both types of attractors are equivalent. This is, however, not generally the case. A counterexample is given by the non-autonomous system

$$x(t+1) = x(t)\exp(-t) \tag{2.46}$$

with the general solution

$$x(t) = x_0 \exp\left(\frac{-(t-1/2)^2 + (t_0 - 1/2)^2}{2}\right) . \tag{2.47}$$

The invariant set $\{0, 0, 0, ...\}$ is a forward attractor of the system, but not a pullback attractor. The opposite is true for $x(t+1) = x(t)\exp(t)$.

For recurrent networks, the notion of pullback and forward attractors will become relevant when introducing echo state networks in Section 3.1. For now, we note that for a recurrent network to reliably perform computations, it is, among other properties, essential that the response to a certain input sequence should be robust against perturbations. This property is captured by the forward attracting property, where the forward attractor consists of a sequence of internal network states.

Similar to the autonomous case, one might be interested in defining a measure that characterizes the properties of the non-autonomous attractor. In analogy to autonomous systems, we can define the largest Lyapunov exponent as the long term growth rate of small perturbations to the system. A practical definition of the largest Lyapunov exponent of a non-autonomous system defined by the flow $\Phi(\boldsymbol{x}, t_0, t)$ can thus be given by

$$l_{\max} \equiv \limsup_{t \to \infty, \|\boldsymbol{\delta}_0\| \to 0} \frac{1}{t - t_0} \ln\left(\frac{\|\boldsymbol{\delta}(t)\|}{\|\boldsymbol{\delta}_0\|}\right) \tag{2.48}$$

$$\boldsymbol{\delta}(t) = \widehat{J}_{t,t_0} \boldsymbol{\delta}_0 \tag{2.49}$$

where $\widehat{J}_{t,t_0}$ is defined in analogy to (2.40) via

$$\widehat{J}_{t,t_0} \equiv \widehat{J}(\Phi(\boldsymbol{x_0}, t_0, t))\, \widehat{J}(\Phi(\boldsymbol{x_0}, t_0, t-1)) ... \widehat{J}(\boldsymbol{x_0}) \tag{2.50}$$

as the evolution of the tangent space [52, 53]. It is interesting to note that the limit superior is used in the definition. This is due to the fact that the explicit time dependence can prevent strict convergence, which is only guaranteed in the autonomous case by the Osedelets theorem [54]. In practice, using (2.48) and (2.49) for numerical purposes can be challenging due to potential divergence issues or when rounding errors occur at the machine precision limit. Different methods have thus been devised for estimating the largest Lyapunov exponent [53, 55] or the entire Lyapunov spectrum [56].

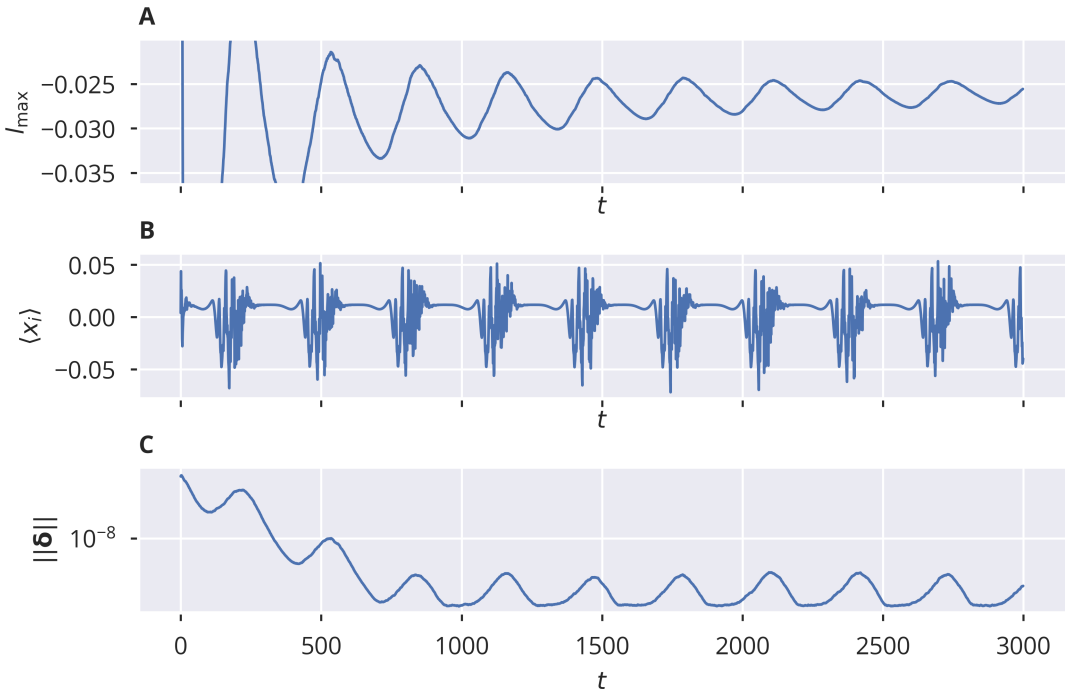

Figure 2.9: Largest Lyapunov exponent of a random recurrent network driven by sinusoidal input. A: Estimate of $l_{\max}$ according to (2.48). B: Average membrane potential. C: Time evolution of the initial perturbation $\boldsymbol{\delta}$, see (2.49).

Fig. 2.9 illustrates the estimation of the largest Lyapunov exponent in a non-autonomous system by means of a network of 1000 randomly connected neurons driven with a sinusoidal input current. The external input causes the network to transition from phases of chaotic, irregular behavior to periods of contracting network dynamics, see Fig. 2.9B. This is reflected in the perturbation vector shown in Fig. 2.9C increasing and decreasing periodically in its length. In the particular case shown here, a general trend towards a decreasing perturbation is observable, which is in line with the estimate of $l_{\max}$ converging towards a value slightly below zero, as shown in Fig. 2.9A.

2.3.3 Mean Field Theory of Large Neural Networks

So far, we have introduced methods and theoretic concepts that are applicable when analyzing small networks, i.e. dynamical systems with relatively low dimensionality. In particular, bifurcation analyses can yield insights into the dynamical behavior of such systems. However, describing the dynamics of recurrent networks of several hundreds or thousands of neurons on such a detailed level can be challenging. For example, since the parameter space of recurrent weights scales quadratic with the number of neurons, finding critical points and characterizing possible bifurcations becomes an almost impossible task for numerical reasons alone. Yet, even a single cubic millimeter of the human visual cortex can contain roughly 3×10^4 to 4×10^4

neurons [57]. Even a cortical minicolumn, which is considered the smallest information processing unit in the cortex, still contains approximately 100 neurons [58, 59]. In addition, due to experimental limitations, a complete picture of the connectivity and dynamical properties of an entire neuronal system is still not available in large mammals, despite massive advancements in recent years [60, 61, 62]

Therefore, similar to systems considered in many-body physics, meaningful predictions can only be made on the level of statistical quantities, describing the expected activity of neurons in a large ensemble. On the other hand, this approach also implies that we should drop the details of a particular synaptic connectome in favor of a statistical distribution over the entire population, allowing for a dramatic reduction in the amount of parameters to be considered. While it might appear at first that extracting meaningful predictions from such a drastic simplification is a hopeless endeavor, one should remind oneself of the fact that each neuron can have thousands of synaptic inputs, making the effect of each single signal relatively small. Therefore, by adding a lot of small input currents, we can expect that the law of large numbers should come into effect, which is a first step towards a statistical understanding of the neuronal dynamics [63].

The Balanced State

Before we progress to a mean-field description of the model at hand, we shall briefly discuss the conditions under which the use of a rate model is justified. When discussing the basics of single neuron models, we stated that using rate model is justified in the case when it is sufficient for modeling the network dynamics. This implies that individual spiking events should occur in irregular, unpredictable patterns. This asynchronous irregular state was first analyzed by van Vreeswijk, Sompolinsky and Brunel [32, 33]. Depending on the details of the network model, the exact mathematical formulation might vary, but the essential condition that should be fulfilled is that the dynamics of the membrane potentials are dominated by fluctuations induced by the superposition of strong excitatory and inhibitory input. For illustrative purposes, we introduce a discrete-time binary spiking network model similar to the one used by van Vreeswijk and Sompolinsky [32]. In the original publication, the model consists of an excitatory and an inhibitory population, with synaptic connections within and across both populations. However, we further simplify the model by not applying the excitatory-inhibitory constraint imposed by Dale's law: We do not assume any topological restrictions with respect to the sign of the synaptic weights. We denote the activity of the i-th neuron in a population time t as $y_i(t)$ and, in the spirit of modeling spiking activity, it is given by the Heaviside function acting on a recurrent input $x_i(t)$, an external input current $I_i(t)$ and a bias b_i:

$$y_i(t) = \Theta\left(x_i(t) + I_i(t) - b_i\right) \ . \tag{2.51}$$

Furthermore, the recurrent input is given by

$$x_i(t) = \sum_{j=1}^{N} W_{ij} y_j(t-1) \; . \tag{2.52}$$

The matrix W_{ij} is usually assumed to be sparse and that $W_{ii} = 0$. To further progress, we introduce a local "rate average" quantity, $r_i(t) = \langle y_i(t) \rangle$, that is defined by van Vreeswijk and Sompolinsky as an average over initial conditions [32, A.1]. In practice, it is the chance of the respective neuron being active at time t averaged over an ensemble of network simulations with different initial activities that are consistent with given initial rates $r_i(0)$. We further assume that the activities $y_i(t)$ are uncorrelated across the population. This does not imply that the population average $\langle y_i(t) \rangle_i$ has to be static over time. Under the assumption of statistical independence, and if the number of afferent presynaptic neurons is large, we can describe $x_i(t)$ as a Gaussian variable with mean $\mu_i(t)$ and variance $\sigma_i^2(t)$ given by

$$\mu_i(t) = \sum_{j=1}^{N} W_{ij} r_j(t-1) \tag{2.53}$$

$$\sigma_i^2(t) = \sum_{j=1}^{N} \mathrm{Var}\left[W_{ij} y_j(t-1)\right] = \sum_{j=1}^{N} W_{ij}^2 r_j(t-1)\left[1 - r_j(t-1)\right] \; . \tag{2.54}$$

For (2.54), we have made use of the fact that for a binary sequence $y_j(t)$, we find $y_j^2(t) = y_j(t)$ and thus $\mathrm{Var}\left[y_j(t)\right] = \langle y_j^2(t) \rangle - \langle y_j(t) \rangle^2 = \langle y_j(t) \rangle - \langle y_j(t) \rangle^2$. The balanced state assumption now poses the condition that the fluctuations in the input should not become arbitrarily small relative to the mean input as the network size increases: The standard deviation, $\sigma_i(t)$, should remain finite. The reason for this assumption is that a finite level of fluctuations guarantees an irregular pattern of spiking activity, which justifies the initial assumption of statistical independence across the population. For simplicity, let us assume a static and symmetric state where $r_i(t) = r$ for all i and t. Then, we find $\mu_i = Nr\langle W_{ij} \rangle_j$ and $\sigma_i^2 = Nr(1 - r)\langle W_{ij}^2 \rangle_j$. For σ_i to stay finite, synaptic weights thus need to scale with $1/\sqrt{N}$. This however, raises the issue that μ_i then scales with $\sqrt{N}$. The only solution here is to impose a "balanced" situation, that is, inhibitory and excitatory cancel each other. This does not imply that we need to find $\langle W_{ij} \rangle_j = 0$ exactly, but excitation and inhibition should keep the mean input on the same order of magnitude as the fluctuations.

Modeling the recurrent input as a Gaussian random variable, we can calculate the expected rate via

$$r_i(t) = \int_{-\infty}^{\infty} \mathcal{N}\left(x, \mu_i(t) + I_i(t) - b_i, \sigma_i^2(t)\right) \Theta(x) \mathrm{d}x \tag{2.55}$$

$$= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu_i(t) + I_i(t) - b_i}{\sqrt{2\sigma_i^2(t)}}\right)\right] \tag{2.56}$$

$$\equiv \psi\left(\frac{\mu_i(t) + I_i(t) - b_i}{\sqrt{2\sigma_i^2(t)}}\right) \tag{2.57}$$

where the function $\mathcal{N}\left(x, \mu, \sigma^2\right)$ denotes a Gaussian density function with mean μ and variance σ^2 and $\mathrm{erf}(x)$ is the error function. As a function of the mean input, $\psi(x)$ has a shape that is very similar to the sigmoidal activation function introduced in (2.7). Yet, the variance $\sigma_i^2(t)$ enters as a quantity modulating the gain. The validity of using a simple sigmoidal type activation function with a static gain and the mean entering as the input therefore depends on how much fluctuations vary over time. There is, of course, no general answer to this: Fluctuations are in turn determined by varying activities within the network, as indicated by (2.54). However, it is also important to consider the fact that (2.54) essentially performs an average over all synaptic afferents, thereby potentially mitigating changes over time.

Ultimately, rate models can only provide some approximation to the dynamics appearing in spiking models. As this simple model illustrates, the predictive power of such models depends on how crucial exact spike times are for predicting the expected activity. Balanced state dynamics justify the treatment of spiking variability as fluctuations entering the membrane potential, leading to sigmoid-like activation functions.

Mean Field Equations for Random Neural Networks

To this point, we did not bother describing the synaptic weights by means of some statistical model other than introducing the general scaling $1/\sqrt{N}$. In this section, we will describe a mean field theory derived from the assumption that weights are independently drawn from a probability density. Early work on the dynamics of recurrent networks with quenched random connectivity have considered autonomous dynamics (i.e. not externally driven) for either continuous [64] or discrete time dynamics [51, 65] and have later been extended to driven recurrent networks [66, 67, 68, 69, 70]. Since our research was concerned with the dynamics of echo state networks, we will introduce key elements of the theory based on Moynot and Samuelides [66] and Massar and Massar [69].

Starting with a network defined by (2.19) and (2.20), we now assume that the network size N is large, W_{ij} is sparse and that non-zero elements are drawn from a continuous probability distribution with zero mean and variance σ^2. As usual, diagonal elements are always zero. For simplicity, even though not strictly necessary,

we can choose a Gaussian distribution. Furthermore, the external weights are also assumed to be drawn independently from a probability distribution with mean μ_{ext} and variance σ^2_{ext}.

To derive a mean-field description of the system, we make the same assumption for the firing rates in the system as we did for the spiking events in the previous section: We assume statistical independence across the population. The recurrent part of the membrane potential

$$x_{\text{r},i}(t) = \sum_{j=1}^{N} W_{ij} y_j(t-1) \tag{2.58}$$

therefore is a sum over statistically independent terms, which means that for $N \to \infty$, the distribution over the population is given by a Gaussian distribution with mean

$$\mu_{\text{x,r}}(t) \equiv \mathrm{E}\left[x_{\text{r},i}(t)\right]_i = N\mathrm{E}\left[W_{ij} y_j(t-1)\right]_{ij} \tag{2.59}$$

and variance

$$\sigma^2_{\text{x,r}}(t) \equiv \mathrm{Var}\left[x_{\text{r},i}(t)\right]_i = N\mathrm{Var}\left[W_{ij} y_j(t-1)\right]_{ij} \ . \tag{2.60}$$

By assumption, weights and activities are also statistically independent, which allows for a factorization, yielding $\mu_{\text{x,r}}(t) = 0$ and

$$\sigma^2_{\text{x,r}}(t) = N\mathrm{E}\left[y_j^2(t-1)\right]_j \mathrm{Var}\left[W_{ij}\right]_{ij} \tag{2.61}$$

$$= N\mathrm{E}\left[y_j^2(t-1)\right]_j \sigma^2 \ . \tag{2.62}$$

For convenience, and in line with the scaling of weights introduced in the previous section on the balanced state, we redefine σ^2 by a parameter g via $\sigma^2 = g^2/N$. For the external input, we can consider two limit cases: First, the input sequence could be scalar valued, that is, a single input stream $y_{\text{ext}}(t)$ is projected onto the neurons via weights $w_{\text{ext},i}$. In this case, the distribution over the population of external inputs at time t is a Gaussian distribution with mean $y_{\text{ext}}(t)\mu_{\text{ext}}$ and variance $y^2_{\text{ext}}(t)\sigma^2_{\text{ext}}$. In the other extreme, the dimensionality of the input could be large enough to argue that each neuron receives a sum of independent input streams, which, again, can be modeled by a Gaussian distribution, except that in this case, the mean is $N_{\text{ext}}\mathrm{E}[y_{\text{ext},i}(t)]_i\mu_{\text{ext}}$ and the variance $N_{\text{ext}}\mathrm{E}[y^2_{\text{ext},i}(t)]_i\sigma^2_{\text{ext}}$. In any case, we denote by $\mu_{\text{x,ext}}(t)$ and $\sigma^2_{\text{x,ext}}(t)$ the sequence of population means and variances of the external input, allowing us to write the population mean and variance of the total membrane potential as

$$\mu_{\text{x}}(t) = \mu_{\text{x,ext}}(t) \tag{2.63}$$

$$\sigma^2_{\text{x}}(t) = g^2\mathrm{E}\left[y_j^2(t-1)\right]_j + \sigma^2_{\text{x,ext}}(t) \tag{2.64}$$

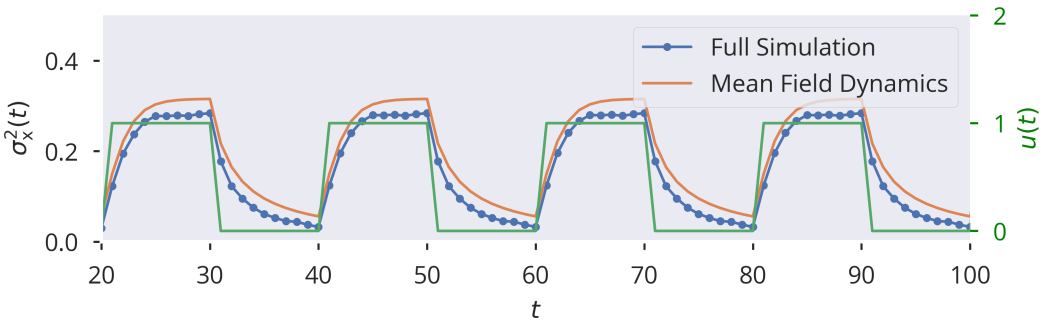

Figure 2.10: Mean field dynamics compared to full simulation in a driven recurrent network. A random neural network with $N = 1000$ and $g = 1$ is driven by the oscillatory binary sequence shown in green. The population variance shows relaxational behavior, correctly predicted by the mean field dynamics as defined in (2.66), using the approximation given by (2.68).

To close the loop, we calculate $\mathrm{E}[y_j^2(t)]_j$ by

$$\mathrm{E}\left[y_j^2(t)\right]_j \equiv \sigma_\mathrm{y}^2\left(\mu_{\mathrm{x,ext}}(t), \sigma_\mathrm{x}^2(t)\right) = \int_{-\infty}^{\infty} \phi^2(x)\, \mathcal{N}\left(x, \mu_{\mathrm{x,ext}}(t) - b, \sigma_\mathrm{x}^2(t)\right) \mathrm{d}x \,. \tag{2.65}$$

Hence, the population mean $\mu_\mathrm{x}(t)$ simply follows the average external input and the time evolution of the population averaged variance of the membrane potential can be determined by evolving a one-dimensional discrete map given by

$$\sigma_\mathrm{x}^2(t) = g^2 \sigma_\mathrm{y}^2\left(\mu_{\mathrm{x,ext}}(t-1), \sigma_\mathrm{x}^2(t-1)\right) + \sigma_{\mathrm{x,ext}}^2(t) \,. \tag{2.66}$$

The integral involved in the function $\sigma_\mathrm{y}^2\left(\mu, \sigma^2\right)$ can not be solved analytically for the usual choice of $\phi(x) = \tanh(x)$. However, we can use an approximation of the hyperbolic function that allows for an analytic solution, which we will repeatedly use for analytic purposes. The approximation reads

$$\tanh(x) \approx \mathrm{sign}(x)\sqrt{1 - \exp\left(-x^2\right)} \tag{2.67}$$

and has a maximum relative error of approximately 4.6%. Using this approximation, the integral can be evaluated to

$$\sigma_\mathrm{y}^2\left(\mu, \sigma^2\right) \approx 1 - \frac{\exp\left(-\mu^2/(1+2\sigma^2)\right)}{\sqrt{1+2\sigma^2}} \tag{2.68}$$

For the example shown in Fig. 2.10, we simulated a sparse random network with 1000 neurons, a weight scaling of $g = 1$ and projected a single binary sequence—switching between an up and down state every 10 time steps—with random Gaussian weights onto the neurons. To predict the dynamics of the variance of the membrane potentials across the population, we used (2.68). The change in the external input shifts the fixed point of σ_x^2, and the network relaxes towards it.

On the mean field level, the Lyapunov exponent of the non-autonomous system (2.66) is generally negative. However, this does not imply stability of the full network. Yet, it is possible to derive an analytic expression for the expected largest Lyapunov exponent of the full system, see Massar and Massar [69]. We define the difference between two initially close solutions $x_i(t)$ and $\tilde{x}_i(t)$ as $\delta_i(t) = x_i(t) - \tilde{x}_i(t)$. Since we are free to choose the initial state, we can, for simplicity, assume that $\delta_i(0) = \delta_0$ for all neurons. Using (2.48), we get

$$l_{\max} = \limsup_{t\to\infty} \frac{1}{2t} \ln\left(\frac{\mathrm{E}\left[\delta_i^2(t)\right]_i}{\delta_0^2}\right) \tag{2.69}$$

and therefore have to determine the evolution of the population averaged perturbation. In general, we find

$$\delta_i(t) = \sum_{j=1}^{N} W_{ij}\phi'\left(x_j(t-1)\right)\delta_j(t-1) \tag{2.70}$$

where $\phi'(x)$ is the first derivative of the activation function. For $\mathrm{E}\left[\delta_i^2(t)\right]_i$, this results in

$$\begin{aligned}
\mathrm{E}\left[\delta_i^2(t)\right]_i &= \sum_{j,k=1}^{N} \mathrm{E}\left[W_{ij}W_{ik}\right]\phi'\left(x_j(t-1)\right)\phi'\left(x_k(t-1)\right)\delta_j(t-1)\delta_k(t-1) && (2.71)\\
&= \sum_{j=1}^{N} \mathrm{E}\left[W_{ij}^2\right]\phi'^2\left(x_j(t-1)\right)\delta_j^2(t-1) && (2.72)\\
&= g^2\mathrm{E}\left[\delta_i^2(t-1)\right]_i \int_{-\infty}^{\infty} \phi'^2(x)\,\mathcal{N}\left(x, \mu_{\mathrm{x,ext}}(t-1) - b, \sigma_{\mathrm{x}}^2(t-1)\right)\mathrm{d}x && (2.73)
\end{aligned}$$

where we assumed the usual statistical independence of presynaptic activities, as well as across the synaptic weights. depending on whether the external input statistics are static or time dependent, the map given by (2.73) can then either be used to directly calculate $l_{\max}$ using the stationary solution of σ_{x}^2 or as an average given by

$$l_{\max} = \ln(g) + \limsup_{t\to\infty} \frac{1}{2t} \sum_{k=0}^{1-t} \ln\left(\tilde{\sigma}_{\mathrm{y}}^2\left(\mu_{\mathrm{x,ext}}(t-1) - b, \sigma_{\mathrm{x}}^2(t-1)\right)\right)\ , \tag{2.74}$$

where we defined by

$$\tilde{\sigma}_{\mathrm{y}}^2\left(\mu_{\mathrm{x,ext}}(t) - b, \sigma_{\mathrm{x}}^2(t)\right) \equiv \int_{-\infty}^{\infty} \phi'^2(x)\,\mathcal{N}\left(x, \mu_{\mathrm{x,ext}}(t) - b, \sigma_{\mathrm{x}}^2(t)\right)\mathrm{d}x \tag{2.75}$$

the population average of the square of the first derivative of the activation function, in correspondence to (2.65).

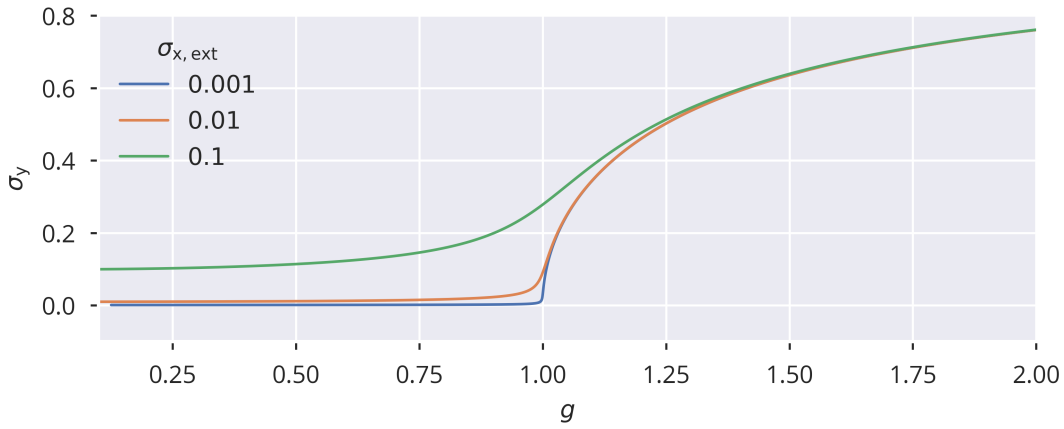

Figure 2.11: Solution of the self-consistent mean field equation (2.76).

For the particular case of stationary external input statistics with zero mean, we can derive from (2.66) and (2.68) a self-consistency equation for σ_y^2:

$$\left(1-\sigma_y^2\right)^2\left(1+2g^2\sigma_y^2+2\sigma_{x,ext}^2\right)=1\,. \tag{2.76}$$

Fig. 2.11 shows solutions for different values of the external standard deviation $\sigma_{x,ext}$. A second order phase transition can be observed at $g=1$. For $\sigma_{x,ext}=0$, this transition corresponds to the onset to chaotic behavior as described by Sompolinsky et al. [64]. In the $\sigma_{x,ext}=0$ case, it can also be explicitly solved for σ_y^2, giving

$$\sigma_y^2=\begin{cases} g\geq 1 & : \frac{4g^2-1-\sqrt{1+8g^2}}{4g^2} \\ g<1 & : 0 \end{cases}\,. \tag{2.77}$$

This solution has a first-order expansion at the critical point $g=1$ of $\sigma^2 y\approxeq 4/3g$ for $g\geq 1$. This means that, close to the critical point $g_c=1$, the standard deviation σ_y shown in Fig.2.11 scales as $(g-g_c)^{1/2}$, corresponding to the critical exponent obtained in the Landau theory for continuous, second-order phase transitions [71, p. 431–432]. If we denote the differential zero-field susceptibility to the external driving $\sigma_{x,ext}$ as $\chi=\partial\sigma_y/\partial\sigma_{ext}|_{\sigma_{ext}=0}$, we obtain the critical scaling $\chi\propto(g_c-g)^{-1/2}$ for $g<g_c$, showing that the responsiveness of the of the system to small external perturbations diverges at the critical point. Note, however, that the obtained critical exponent $1/2$ is not the same as one would expect for the susceptibility in the Landau theory for second-order phase transitions, which is 1.

Non-zero external driving smoothes out the transition, which also has the effect of stabilizing the network, i.e. reducing l_{max} [72, 69]. This effect can be explained by the second term appearing in (2.74): External input tends to push the membrane potential away from zero. Due to the sigmoidal shape of the activation function, this causes the average over the squared derivative $\phi'^2(x)$ to become less than one, making the entire term negative, decreasing l_{max}.

2.4 Neuronal Homeostasis

So far, we described neuronal dynamics as a result of a rigid architecture defined by synaptic weights as well as intrinsic parameters like gains and biases. This static perspective is, evidently, an idealization. In reality, an abundance of mechanisms constantly alter intrinsic and synaptic properties of neural networks [73, 74, 8, 75, 76, 77]. Some of these mechanisms have been summarized under the notion of *neural homeostasis*. The term *homeostasis* was introduced by Walter Cannon [78, 79] and refers to the capability of a biological system to approach and maintain a stable internal state under external physical influences. A well known example is the ability of mammals to keep their core body temperature relatively constant in the face of changing environment temperatures.

Similar effects have been observed in the brain, where neuronal activity remains quite stable given that the brain is subject to structural changes during development, as well as changes in external stimuli. Therefore, it is generally acknowledged that neuronal activity must be controlled by some form feedback mechanism. Indeed, many different types of regulatory processes have been experimentally confirmed, acting on synaptic efficacies as well as intrinsic neuronal parameters [7, 75, 80].

On the theoretic side, feedback dynamics generally fall into the field of control theory [81]. In control systems, the target quantity that is to attained is usually called a *set point* and the system attempts to minimize the distance between some measure of its internal state and this predefined quantity. For neuronal homeostasis, this quantity is usually assumed to be the average firing rate. Apart from metabolic advantages [82], operating in a certain activity range is also assumed to be beneficial for computational capabilities [83, 9]. We will further discuss the unsupervised optimization of hyperparameters for echo state networks in Section 3.1.

2.4.1 Dual Homeostasis

Given the fact that numerous realizations of such control mechanisms can exist, the question arises how these different processes work together in controlling neural activity. In the worst-case scenario, two or more control mechanisms conflicting in their individual set points might cause instabilities, not only failing to reach any of the homeostatic targets but causing the state of the system to undergo fast, chaotic and potentially harmful transitions [10]. Here, we introduce a theoretic treatment of the most simple polyhomeostatic case, i.e. two parameters of the neuronal system acting upon the neuronal dynamics via two separate feedback loops. This is largely based upon the work of Cannon and Miller [11] on dual homeostasis.

Suppose two local neuronal control parameters, a and b, affect some intrinsic, time dependent physical quantity $x(t, a, b)$ that is to be regulated. As an example, x could be the internal membrane potential, but also some estimate of the firing rate. In the most general sense, a homeostatic mechanism acting on a and b could

be written as

$$\Delta a(t) = \epsilon_{\mathrm{a}} f\left(x\left(t, a(t), b(t)\right), a(t), b(t)\right) \tag{2.78}$$

$$\Delta b(t) = \epsilon_{\mathrm{b}} g\left(x\left(t, a(t), b(t)\right), a(t), b(t)\right) \tag{2.79}$$

where $a(t+1) = \Delta a(t) + a(t)$ and $b(t+1) = \Delta b(t) + b(t)$ describe discrete time dynamics. Since we assume both adaptation rates ϵ_{a} and ϵ_{b} to be small, the same expressions could also be used for $\dot{a}$ and $\dot{b}$ in continuous time models. Note that the homeostatic feedback functions f and g are dependent on the variable x that is to be controlled, but can also have some explicit dependence on the control parameters a and b. An example for the latter would be the case where the rate of change is proportional to the control parameter itself, meaning e.g. $f(x, a, b) = a\tilde{f}(x)$, which prevents a from becoming negative. As long as x is not static, we can, in general, not expect to find a fixed point of the dynamics in the strict sense, that is $\Delta a(t) = \Delta b(t) = 0\ \forall t$. However, if we assume that adaptation is slow compared to the time scale of the fluctuations in x and that x is a stationary sequence, we can define a weaker form of fixed point by means of the temporal average as $\langle \Delta a(t) \rangle_t = \langle \Delta b(t) \rangle_t = 0$. More generally, we can state that the non-autonomous adaptation dynamics (2.78) and (2.79) can be approximated by the stationary system

$$\langle \Delta a(t) \rangle_t = \epsilon_{\mathrm{a}} \left\langle f\left(x\left(t, a, b\right), a, b\right) \right\rangle_t \tag{2.80}$$

$$\langle \Delta b(t) \rangle_t = \epsilon_{\mathrm{b}} \left\langle g\left(x\left(t, a, b\right), a, b\right) \right\rangle_t \ . \tag{2.81}$$

While this generic form does not allow for much further analysis, it should already become apparent that such a system is not guaranteed to have a fixed point: Since a and b are coupled via their effect on x, the function f could drive a into a direction that would make it impossible for b to reach a stationary state, and vice versa. For further analysis, Cannon and Miller [11] make the assumption that both f and g are well described by a quadratic equation over the support of the probability distribution of x. Since we also included the possibility of an explicit dependence on a and b, we generalize this assumption to the case where f and g are still approximately quadratic in x, but the coefficients are functions of a and b:

$$\epsilon_{\mathrm{a}}^{-1} \langle \Delta a \rangle \approx f_0(a,b) + f_1(a,b)\langle x \rangle + \frac{1}{2} f_2(a,b)\langle x^2 \rangle \tag{2.82}$$

$$\epsilon_{\mathrm{b}}^{-1} \langle \Delta b \rangle \approx g_0(a,b) + g_1(a,b)\langle x \rangle + \frac{1}{2} g_2(a,b)\langle x^2 \rangle \ . \tag{2.83}$$

For simplicity, we have dropped the explicit notation of temporal averages and the dependence of x on a and b, which is still present. Setting the left hand side to zero, one finds a linear system for the fixed point of the adaptation, which can be solved

for the first and second moment of x, resulting in

$$\langle x \rangle = \frac{f_2 g_0 - g_2 f_0}{f_1 g_2 - g_1 f_2} \tag{2.84}$$

$$\langle x^2 \rangle = 2\frac{g_1 f_0 - f_1 g_0}{f_1 g_2 - g_1 f_2} \, . \tag{2.85}$$

Therefore, *if* a steady state solution of the adaptation exists, it determines the mean and variance of the variable x that is to be controlled. Furthermore, a unique solution can only be present if the denominator $f_1 g_2 - g_1 f_2$ is not zero. In particular, the denominator becomes zero if the dynamics of both a and b are only linear (meaning $f_2 = g_2 = 0$). Moreover, the fact that a solution exists does not imply that it can be attained by a certain a and b. For example, obviously, a negative $\langle x^2 \rangle$ predicted by (2.85) is not a viable solution.

If the coefficients entering (2.84) and (2.85) are independent of a and b, the corresponding fixed point (a^*, b^*) can be directly determined by finding a pair (a, b) resulting in the correct first and second moment of x. In general, however, the coefficients in (2.84) and (2.85) are dependent on a and b and the solution might not be readily available.

Stability

If a fixed point in the dual homeostatic system exists, the next task is to determine whether it is stable under perturbations. For this purpose, we have to evaluate the stability of the linearized system

$$\widehat{J} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}a}\langle f \rangle & \frac{\mathrm{d}}{\mathrm{d}b}\langle f \rangle \\ \frac{\mathrm{d}}{\mathrm{d}a}\langle g \rangle & \frac{\mathrm{d}}{\mathrm{d}b}\langle g \rangle \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\partial f}{\partial x}\frac{\partial x}{\partial a} \right\rangle + \left\langle \frac{\partial f}{\partial a} \right\rangle & \left\langle \frac{\partial f}{\partial x}\frac{\partial x}{\partial b} \right\rangle + \left\langle \frac{\partial f}{\partial b} \right\rangle \\ \left\langle \frac{\partial g}{\partial x}\frac{\partial x}{\partial a} \right\rangle + \left\langle \frac{\partial g}{\partial a} \right\rangle & \left\langle \frac{\partial g}{\partial x}\frac{\partial x}{\partial b} \right\rangle + \left\langle \frac{\partial g}{\partial b} \right\rangle \end{pmatrix} \tag{2.86}$$

at the fixed point.

A Simple Example

For illustrative purposes, we can consider the a dual-homeostatic system on the neuronal firing rate, given by $y = \tanh(ax - b)$ and some input $x(t)$ that is a random Gaussian variable with zero mean and unit variance. For the feedback dynamics we choose

$$\epsilon_{\mathrm{a}}^{-1} \langle \Delta a \rangle = \langle y \rangle - \langle y^2 \rangle \tag{2.87}$$

$$\epsilon_{\mathrm{b}}^{-1} \langle \Delta b \rangle = -\frac{1}{2} + \langle y \rangle \, . \tag{2.88}$$

According to (2.84) and (2.85), we should expect for a fixed point to yield both $\langle y^* \rangle = 0.5$ and $\langle y^{2^*} \rangle = 0.5$.

In Fig. 2.12, we show both the resulting dynamics on a and b as well as the corresponding dynamics on the first and second moment of y. As correctly predicted,

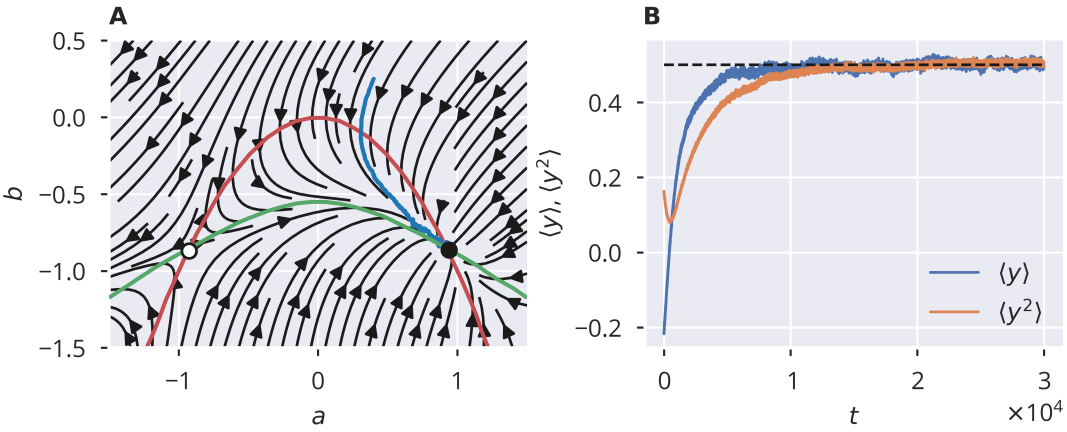

Figure 2.12: Homeostatic dynamics given by (2.87) and (2.88) and $y = \tanh(ax - b)$, where x is a standard Gaussian random variable with zero mean and unit variance. A: Flow of the system. Shown in red and green are the nullclines of a and b, respectively. The white dot denotes a saddle point, the black point a stable fixed point. The blue line is the trajectory of a simulation serving as an example. B: Dynamics of the first two moments of neuronal activity corresponding to the blue line in A. As predicted, both moments settle to 0.5 (dashed line).

both $\langle y \rangle$ and $\langle y^2 \rangle$ settle at 0.5. However, the flow of the system shown in Fig. 2.12A reveals that two fixed points exist, yet one is a saddle point. Thus, while the homeostatic control is locally stable against perturbations around the stable fixed point, it is not globally stable.

Controlling the Membrane Potential by Firing Rate Homeostasis

In the previous example, the average squared activity enters the adaptation dynamics for the scaling factor a. While not entirely implausible, one might be tempted to consider it biologically questionable to include higher orders of neuronal firing rates. Yet, for controlling the first and second moment of y, this was inevitable. Here, we would like to give another example of dual homeostasis that only uses the first moments of physically interpretable quantities, that is, the effective membrane potential $\tilde{x} = ax - b$ and the firing rate y. We choose the dual homeostatic mechanism to simply be

$$\epsilon_{\mathrm{a}}^{-1} \langle \Delta a \rangle = \langle \phi(\tilde{x}) \rangle - \mu_{\mathrm{y}} \tag{2.89}$$

$$\epsilon_{\mathrm{b}}^{-1} \langle \Delta b \rangle = \langle \tilde{x} \rangle - \mu_{\tilde{\mathrm{x}}} \,. \tag{2.90}$$

Furthermore, we choose our activation function to be $\phi(x) = [1 + \mathrm{erf}(2x)]/2$ in this case, which has two advantages: First, it is a continuous, strictly positive nonlinear function within $[0, 1]$, which makes it simpler to interpret at as a firing rate. Second, the use of the error function instead of the hyperbolic function allows us to analytically determine the fixed points resulting from the dynamics.

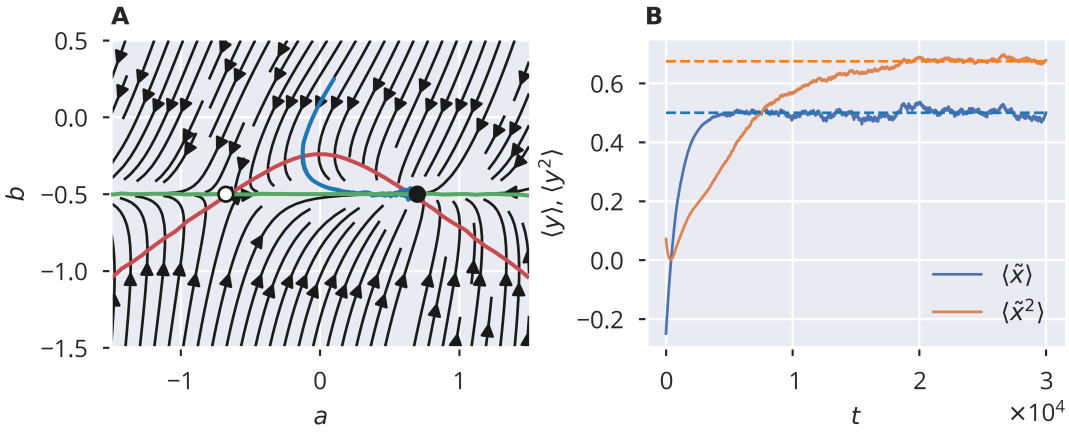

Figure 2.13: Homeostatic dynamics given by (2.89) and (2.90). A: Flow of the system, using the same scheme as in Fig. 2.12. B: Dynamics of the first two moments of the effective membrane potential $\tilde{x} = ax - b$ for the blue trajectory in A. The dashed lines mark the corresponding analytic predictions.

First, we note that the fixed points per se are attained if $\langle \tilde{x} \rangle = \mu_{\tilde{\mathrm{x}}}$ and $\langle y \rangle = \mu_{\mathrm{y}}$, meaning that this control directly regulates the average firing rate as well as the average internal effective membrane potential. Since we assume, as in the previous example, that the input signal x has zero mean and unit variance, we directly find $b^* = -\mu_{\tilde{\mathrm{x}}}$ for the fixed point coordinate of b. For Gaussian input, the average of the firing rate in (2.89) can be explicitly evaluated by a Gaussian integral, resulting in

$$\langle \phi(\tilde{x}) \rangle = \frac{1}{2} \int_{-\infty}^{\infty} \frac{\exp(-x^2)}{\sqrt{2\pi}} \left[1 + \operatorname{erf}\left(2(ax - b)\right)\right] \mathrm{d}x \tag{2.91}$$

$$= \frac{1}{2} \left[1 - \operatorname{erf}\left(\frac{b}{\sqrt{2a^2 + 1/4}}\right)\right] . \tag{2.92}$$

Since we already found b^*, we can use this explicit expression to solve for a^*, giving

$$a^* = \pm \sqrt{\frac{\mu_{\tilde{\mathrm{x}}}^2}{2\left[\operatorname{erf}^{-1}\left(1 - 2\mu_{\mathrm{y}}\right)\right]^2} - \frac{1}{8}} . \tag{2.93}$$

Furthermore, we can calculate the second moment of $\tilde{x}$ at the fixed point from the solution of a^* via $\langle \tilde{x}^2 \rangle = a^{*2} + \mu_{\tilde{\mathrm{x}}}^2$. Thus, the homeostatic control has fixed points each associated with some first two moments of $\tilde{x}$. However, it does not explicitly require $\tilde{x}^2$ or higher moments to enter the dynamics as it utilizes the inherent nonlinearity of the activation function ϕ. In Fig. 2.13, the dynamics of the system are shown for $\mu_{\mathrm{y}} = 3/4$ and $\mu_{\tilde{\mathrm{x}}} = 1/2$. The flow is quite similar to the previous example in that one of the two fixed points is stable and one is a saddle node.

This example goes to show that stable dual homeostasis does not necessarily imply that higher moments of physical quantities have to contribute to the dynamics of the adaptation if nonlinear relationships that are already inherent to the system can be employed. However, even though the particular choice of activation function in the example allowed us to analytically determine the fixed points of the system, in general, predicting fixed points of the adaptation for arbitrary nonlinear systems can of course potentially be more difficult compared to the simple quadratic form introduced earlier. Still, if the fluctuations of the driven system only weakly couple to third order or higher contributions in the adaptation, a second-order Taylor expansion might give good predictions regarding the steady state of the adaptive system.

CHAPTER 3

Flow Control

Fabian Schubert, Claudius Gros
Local Homeostatic Regulation of the Spectral Radius of Echo-State Networks,
Front. Comput. Neurosci. (2021)
[84]

In Section 2.4, we introduced neuronal homeostasis as a means to regulate activity in neural networks. A particular variant, synaptic scaling, has been successfully applied in recurrent neural networks [85, 86, 87, 88, 89]. Still, these works used synaptic scaling either as the only homeostatic control [87, 86], or resorted to a form of multiplicative normalization of weights [85, 88, 89]. Therefore, controlling higher moments of neuronal activity is not possible in these control schemes, see Section 2.4.1 on dual homeostasis.

On a more abstract level, a combination of intrinsic homeostasis and synaptic scaling can be mapped to an adjustment of the bias and gain of a nonlinear activation function. A combined adaptation of these parameters has been investigated with respect to their effect on the computational capabilities of networks. A common practice for deriving appropriate learning rules is to define a target probability distribution for the neuronal activity rather than a single set point. The Kullback-Leibler divergence between the target distribution and the empirical distribution can then be minimized via gradient descent [9, 90, 91, 92]. Alternative approaches were also based on information-theoretic measures [83].

These studies did indeed find computational advantages if the target distribution was appropriately chosen. On the other hand, other studies suggested that the spectral radius of the recurrent weight matrix, see Section 2.3.2, is a crucial parameter for optimization [91]. In consequence, while optimizing a network with respect to a given distribution appears to be a good strategy from a machine learning point of view, from a mechanistic, biological perspective, it is likely that an additional control mechanism is required regulating the overall scaling of the recurrent weights, or the spectral radius, as a representative measure.

To the best of our knowledge, theoretical work on spiking neural networks did not investigate the effect of the spectral radius on network dynamics and task performance. However, as introduced in Section 2.3.3, balanced state dynamics in recurrent spiking networks imply a scaling of synaptic weights as a function of the number of

presynaptic connections k by $1/\sqrt{k}$ (which is equivalent to $1/\sqrt{N}$ where N is the number of neurons if the network is densely connected). It follows from the circular law for large random matrices [93] that this scaling keeps the spectral radius finite as the network size and the number of afferent connections increases. Indeed, experiments have verified this scaling [94].

In [84], we explored the possibility of controlling the spectral radius of a recurrent network by combining intrinsic adaptation with a local synaptic scaling rule. For this purpose, we used an echo state framework [95, 96], which we introduce in Section 3.1. While echo state networks as a form of reservoir computing are mostly known in the neuro-inspired machine learning community, they can also be regarded as an abstract representation of neuronal dynamics and computation in recurrent cortical networks [97, 98, 99]. From another perspective, controlling the spectral radius would allow to tune the dynamics of the network to the "edge of chaos" [100], which is considered beneficial for information processing.

The major challenge of such a local learning rule is the non-locality of the spectral radius: It is a quantity defined by all synaptic weights. Yet, individual neurons only have access to a limited amount of information about the entire network state. More precisely, a neuron only has access to its own internal physical state and the presynaptic activity arriving at its afferent synapses, and any local adaptation mechanism is necessarily restricted to utilizing those quantities. In a less restrictive sense, one could also argue that it should be possible to have access to some population- averaged quantities, such as information about neighboring neuronal activity transmitted by diffusive neurotransmitters [101].

In this chapter, we will present an unsupervised homeostatic mechanism, termed *flow control* that is local in its variables but is capable of regulating the spectral radius of the network, thus providing a solution to the described issue of controlling the spectral radius as a non-local quantity. Functionally, it regulates the mean and variance of neuronal activity such that the network operates in a state suitable for sequence learning tasks. As indicated, it affects both biases and the scaling of recurrent synaptic weights.

After a general description of the echo state framework, we will present the adaptation rules of flow control and discuss their capacity of controlling the spectral radius. This is followed by an evaluation of the sequence processing performance of networks that were subject to homeostatic adaptation.

3.1 Echo State Networks

Inspired by the highly recurrent networks found in the brain, recurrent neural network models have proven to be a powerful tool in processing sequential information, due to their combination of dynamic memory and nonlinear processing capabilities. While their complex dynamics are a strength of these networks, it also poses a problem which, in reverse, still concerns research on biological networks: How do recurrent networks learn?

In contrast to computational models of biological networks, utilizing neural network models for machine learning purposes does not impose any biological constraints on potential learning algorithms. One of the most prominent solutions for supervised learning applications, error backpropagation (BP) [102], was initially invented for non-recurrent networks and relies on the propagation of training errors through the network. The ideas was transferred to recurrent networks shortly after and is now known as backpropagation through time (BPTT) [103, 104]. Apart from exhibiting practical challenges such as potential instabilities in the learning algorithm that prevent convergence of the training process, it appeared unlikely that such a learning algorithm could resemble learning processes in the brain. Therefore, while backpropagation did inspire research aiming to transfer the framework to hierarchical, layered networks in biology [13, 15, 105], there is no evidence that temporal learning in the brain could utilize a principle similar to BPTT.

In the last two decades, reservoir computing has emerged as an alternative approach to training recurrent networks [96, 106, 107]. Reservoir computing, as an overarching principle, subsumes methods for sequential information processing that utilize high-dimensional, non-linear dynamical systems as "dynamic reservoirs" which are, in some way, coupled to the sequential input that is to be processed. Essential to the idea is that the learning process does not affect the internal architecture or parameters of the dynamical system. Rather, the output that is to be generated is a simple linear projection from the high-dimensional internal state of the system. Echo state networks are a particular realization of the concept, using large, sparse random recurrent networks with rate encoding neurons as the dynamical reservoir [95, 96, 108, 72]. On the practical side, this idea radically simplifies the learning process: It reduces to a simple fitting of a linear model. On the biological side, it offers a potential solution to the constraint that learning should be local: Readout neurons, see Fig. 2.5, can calculate a linear superposition of the recurrent network activity via their synaptic afferents and adjust synaptic efficacies by some local plasticity rule that utilizes information from additional inputs encoding error signals or targets. The necessity of passing learning errors back into the recurrent network is therefore completely eliminated. Indeed, the dynamical behavior found in neuronal reservoir models has also been identified in cortical networks [99].

In the following, we introduce a mathematical definition of the echo state architecture, much of which has already been introduced in Sections 2.3 and 2.3.3. The essential components of an echo state network are contained in the equations given

in (2.19)–(2.21):

$$\boldsymbol{y}(t) = \phi\left(\boldsymbol{x}(t) - \boldsymbol{b}\right) \tag{3.1}$$

$$\boldsymbol{x}(t) = \widehat{W}\boldsymbol{y}(t-1) + \widehat{W}_{\text{ext}}\boldsymbol{y}_{\text{ext}}(t-1) \tag{3.2}$$

$$\boldsymbol{x}_{\text{out}}(t) = \widehat{W}_{\text{out}}\boldsymbol{y}(t)\ . \tag{3.3}$$

As introduced before, echo state networks traditionally use tanh-activation functions. Furthermore, the network is typically large (e.g. $N = 1000$), but sparsely and randomly connected, except for self-connections which are always absent. Connection probabilities between nodes in the network are usually chosen to be in the low percentage range (we usually chose a connection probability of $p = 0.1$). Alternatively, one can define a fixed number of afferent connections k that is kept constant for different network sizes. This has the particular computational advantage that the number of computations needed for calculating recurrent inputs only scales linear with the network size, in contrast to $\mathcal{O}\left(N^2\right)$ for dense networks. Non-zero connections are randomly generated from a probability distribution with zero mean, which, in our case, was a Gaussian distribution. Input weights are usually dense and, for our implementation, drawn independently from a Gaussian distribution.

Training of the network is only taking place at the readout weights $\widehat{W}_{\text{out}}$. If the network is supposed to generate a target output sequence $\boldsymbol{f}(t)$, the weights should be the solution to a simple linear regression problem, which usually uses the mean squared error:

$$\widehat{W}_{\text{out}} = \arg\min_{\widehat{V}} \sum_{t=0}^{T-1} \left\| \widehat{V}\boldsymbol{y}(t) - \boldsymbol{f}(t) \right\|^2\ . \tag{3.4}$$

If we denote by $Y_{ij} = y_i(j)$ the matrix whose columns are the neuronal activity vectors for different times, and, likewise, $F_{ij} = f_i(j)$, then the solution to the minimization problem is given by

$$\widehat{W}_{\text{out}} = \widehat{F}\widehat{Y}^T\left(\widehat{Y}\widehat{Y}^T\right)^{-1}\ . \tag{3.5}$$

Optionally, we can regularize the objective function by a matrix norm penalizing large readout weights:

$$\widehat{W}_{\text{out}} = \arg\min_{\widehat{V}} \sum_{t=0}^{T-1} \left\| \widehat{V}\boldsymbol{y}(t) - \boldsymbol{f}(t) \right\|^2 + \gamma \left\| \widehat{V} \right\|_F^2\ , \tag{3.6}$$

where $\|\cdot\|_F$ is the Euclidean, or Frobenius matrix norm. In this case, we find

$$\widehat{W}_{\text{out}} = \widehat{F}\widehat{Y}^T\left(\widehat{Y}\widehat{Y}^T + \gamma\widehat{\mathbb{1}}\right)^{-1}\ , \tag{3.7}$$

where $\widehat{\mathbb{1}}$ is the identity matrix. As expected, (3.5) is restored for $\gamma = 0$. For a derivation of (3.7), see Appendix B.1.

From a more biological perspective, learning can also take place in an incremental in-time procedure, for example by using a simple local gradient descent rule, which would read

$$\Delta W_{\text{out},ij} = -\epsilon \left[y_j(t) \left(x_{\text{out},i}(t) - f_i(t) \right) + \gamma W_{\text{out},ij} \right] \ . \tag{3.8}$$

For testing the network performance, we always use (3.7), with the additional choice of adding a single neuron projecting to the readouts which was always fully active, that is $y_{N+1}(t) = 1$ for all t. This is equivalent to adding a bias term to (3.3).

Hyperparameters of Echo State Networks

The echo state framework vastly reduces the number of parameters that have to be optimized for learning a specific task. It comes at the cost, however, of choosing a number of hyperparameters that characterize the statistics of the neuronal reservoir. Even if we restrict ourselves to input and recurrent weights with zero mean and a Gaussian distribution, we still need to choose variances for those distributions, apart from the parameter p that we already introduced. In the following, we will denote the variance of $\widehat{W}_{\text{in}}$ by σ^2_{ext} and the variance of $\widehat{W}$ by σ^2_{w}. Importantly, the latter can be related to the spectral radius $\rho(\widehat{W})$ of the matrix for large N by means of the circular law of random matrices [93]. It states that the spectral radius of a random $N \times N$ matrix with zero mean and variance $1/N$ converges to 1 as $N \to \infty$. In the light of this relation, it makes sense to use the same parameterization of σ^2_{w} as done in Section 2.3.3, $\sigma^2_{\text{w}} = g^2/N$, such that we have $\rho(\widehat{W}) \cong g$ for large N.

Both σ^2_{ext} and g can significantly affect the performance of the network [95, 108]. At the very least, parameters must be chosen such that the network fulfills what is known as the *echo state property* (ESP), which we will discuss in the following.

The Echo State Property

A formal definition of the echo state property can be given by the non-autonomous flow $\Phi(\boldsymbol{x}_0, t_0, t)_u$, see Section 2.3.2, associated with a given echo state network and a left-infinite external input sequence $\mathcal{U} = \{..., \boldsymbol{u}(t-1), \boldsymbol{u}(t)\}$ [95]:

A driven recurrent network with the flow $\Phi(\boldsymbol{x}_0, t_0, t)_u$ generated by a left-infinite input sequence $\mathcal{U} = \{..., \boldsymbol{u}(t-1), \boldsymbol{u}(t)\}$ has the *echo state property* if for all such possible sequences $\mathcal{U}$, the following two conditions are met. (1): unique states $\boldsymbol{x}(\mathcal{U})$, meaning that $\boldsymbol{x}(\mathcal{U}) \neq \boldsymbol{x}(\mathcal{U}')$ if $\mathcal{U} \neq \mathcal{U}'$. (2): $\boldsymbol{x}(\mathcal{U}) = \lim_{t_0 \to -\infty} \Phi(\boldsymbol{x}_0, t_0, t)_u$ for all initial conditions $\boldsymbol{x}_0$.

This definition is very similar to the definition of a pullback attractor as introduced in Section 2.3.2. This resemblance comes from the idea that the effect of differences in the initial internal state of the network should gradually vanish. Moreover, on top of this general condition, the attracting invariant set should be a sequence of individual points in phase space. Finally, each input sequence up to a time t should yield a *unique* state in phase space. The latter condition is not a necessary condition for a pullback attractor (different input sequences could result in the same attractor),

but is essential to the information processing capabilities of the network: If the uniqueness condition is met, the mapping $\boldsymbol{x}(\mathcal{U})$ of all left-infinite time series to the state space $\boldsymbol{x}$ is bijective. This means that, in principle, any function—that is, computation—acting on some entire input sequence could be transformed to a function acting only on the network state $\boldsymbol{x}(t)$ at the last time step. In the basic echo state model, this function happens to be a simple linear transformation given by $\widehat{W}_{\text{out}}$. Other approaches using non-linear models for generating the readout have also been successfully used [109, 110].

The conditions under which the echo state property can be found have been successively refined in recent years [111, 112, 113]. A necessary condition for the echo state property is given by $\rho(\widehat{W}) < 1$. This is relatively easy to see by considering the null-sequence $u(t) = 0,\ \forall t$. If the spectral radius is greater than one, the network enters the chaotic phase, see Fig. 2.11. In this regime, the network is highly sensitive to the initial conditions, which violates the first condition of the ESP. While not being a sufficient condition, in practice, $\rho(\widehat{W}) < 1$ guarantees the ESP [95, 108]. A more strict and sufficient condition is $\sigma_{\max}(\widehat{W}) < 1$, where $\sigma_{\max}$ is the largest singular value. For example, for a zero-mean large random matrix with a largest singular value of 1, the expected spectral radius would be 1/2 [114], which is significantly smaller than the practical condition $\rho(\widehat{W}) < 1$.

Optimization of Hyperparameters

While the ESP is the most basic condition for an ESN to perform sequential processing, it does not guarantee good performance, and many parameter choices that fulfill the ESP can still lead to bad results. Therefore, finding some objective measures helping to optimize the reservoir for better performance has been in issue of ongoing research. One approach is based upon local information-theoretic measures, which was inspired by biological neurons: An exponential distribution of (non-negative) firing rates maximizes entropy for a given mean, and a corresponding intrinsic plasticity rule was first investigated by Triesch [9]. An application to neuronal reservoirs did indeed show an improved performance [90, 115].

Another more general concept is known as the edge-of-chaos hypothesis, stating that the computational capabilities of dynamical systems are improved if they operate close to, but below a transition to chaotic behavior [116, 117, 118, 91, 119]. For small input amplitudes, this transition appears at $\rho(\widehat{W}) = 1$, and the traditional approach to constructing echo state reservoirs was to set the spectral radius to values close to, but below 1 [108]. For stronger inputs, the control parameter g can be pushed to higher values and still remain in the non-chaotic regime [120, 68]. Still, experiments on different sequential tasks showed that adjusting the spectral radius to a value close to one gave generally good performance, comparable to what was achieved when the network was optimized using intrinsic plasticity [121].

An important property of recurrent networks that is directly affected by the spectral radius is the memory capacity of the network. For a given scalar valued input sequence $u(t)$ and a time delay τ, one can define the target output to be a

delayed version of the input: $f_\tau(t) = u(t-\tau)$. The memory capacity of the network is then given by

$$\mathrm{MC} \equiv \sum_{k=0}^{\infty} \rho^2 \left(f_k(t), x_{\mathrm{out}}(t)\right) \ , \tag{3.9}$$

where $\rho^2 \left(f_k(t), x_{\mathrm{out}}(t)\right)$ is the squared Pearson correlation between the target sequence and the generated output. Importantly, the readout weights are independently optimized for each time delay τ. In practice, it is neither possible nor necessary to evaluate an infinite number of terms of the sum, since the maximally achievable delay that can be recalled is bounded by the network size N [122]. For random sequences and linear echo state networks ($\phi(x) = x$), this upper bound is achieved for a spectral radius $\rho(\widehat{W}) \to 1$ from below [122]. For the usual nonlinear activation, the maximally achievable memory capacity is still attained at $\rho(\widehat{W}) = 1$, albeit being smaller than in the linear case.

These results illustrate that a spectral radius close to one is beneficial in cases where extended temporal memory is required. Therefore, optimal tuning of parameters depends on the task at hand. However, from a biological point of view, tuning such hyperparameters should rather be viewed as a form of homeostasis that regulates the recurrent network towards a dynamic regime that is *generically* beneficial for processing temporal information, rather than being a fine-tuning process, tailored for a very specific, single task. Therefore, if we assume that memory is always required to some extent when temporal information is to be extracted from sequential data, tuning the spectral radius towards a value close to one is a reasonable assumption.

3.2 Homeostatic Model

On top of the standard echo state model, we introduce two adaptation rules, affecting recurrent synaptic weights and biases.

First, biases are subject to a simple adaptation rule controlling the mean neuronal activity:

$$b_i(t) = b_i(t-1) + \epsilon_{\mathrm{b}} \left[y_i(t) - \mu_{\mathrm{t}}\right] \ . \tag{3.10}$$

Here, ϵ_{b} controls the adaptation rate and μ_{t} is the target activity.

For the synaptic scaling rule, we added a local multiplicative factor, $a_i(t)$, which defines an effective synaptic weight matrix $W_{\mathrm{a},ij}(t) \equiv a_i(t) W_{ij}$ with a spectral radius R_{a} that is to be controlled. Accordingly, the recurrent contribution to the membrane potential is then given by

$$x_{\mathrm{r},i} = a_i(t) \sum_{j=1}^{N} W_{ij} y_j(t-1) \ . \tag{3.11}$$

The multiplicative factor $a_i(t)$ is subject to an adaptation rule that we termed *flow control*:

$$a_i(t) = a_i(t-1)\Big[1 + \epsilon_\mathrm{a}\Delta R_i(t)\Big], \qquad \Delta R_i(t) = R_\mathrm{t}^2 y_i^2(t-1) - x_{\mathrm{r},i}^2(t)\ . \tag{3.12}$$

Here, R_t is the target spectral radius that is to be achieved, and ϵ_a is small factor controlling the adaptation rate. We also define an alternative non-local rule that we used mainly for comparative purposes when testing the effectiveness of (3.12)., Here, the change $\Delta R_i(t)$ reads

$$\Delta R_i(t) = \frac{1}{N}\Big[R_\mathrm{t}^2\,||\boldsymbol{y}(t-1)||^2 - ||\boldsymbol{x}_\mathrm{r}(t)||^2\Big]\ . \tag{3.13}$$

A list of the used parameter values is given in Table 3.1. Note that adaptation is generally slow, since its dynamics should capture temporal averages over the fluctuating quantities entering (3.10) and (3.12).

For (3.12) to be biologically plausible, we had to make a number of assumptions. First, the only the recurrent part of the synaptic input enters the equation. This means that we had to assume some physical separation between recurrent connections and external input. A possible justification for this approach is the anatomical structure of pyramidal neurons, which would allow for a physical separation of different input types [123]. In return, the scaling also only affects recurrent weights and not external weights. This separation is, however, not essential for the successful tuning of the spectral radius, since external weights have no effect on $\rho(\widehat{W})$. An additional assumption that we had to make was to allow second moments of y and x_t to enter the adaptation rule. While it might appear questionable that such quantities are physically represented in the cell, we have illustrated in the second example given in 2.4.1 that the second moment of physical quantities can be controlled via polyhomeostatic control if intrinsic nonlinearities between the components of the system are utilized. Similar models of biologically inspired intrinsic plasticity have also been proposed using complex functions of firing rates and membrane potentials [9, 90, 121].

3.2.1 Theoretical Motivation

As described in the introduction to echo state networks, the circular law for random matrices states that for large N, the eigenvalues are distributed uniformly on the complex unit disc if the variance of the underlying distribution is $1/N$ [93]. Therefore, if the synaptic scaling factors of a large randomly initialized recurrent weight matrix with entries W_{ij} were uniformly set to $1/\sqrt{N\sigma_\mathrm{w}^2}$, where σ_w is the variance

Table 3.1: Standard values for model parameters

N	p_r	μ_t	ϵ_b	ϵ_a
500	0.1	0.05	10^{-3}	10^{-3}

of the entries W_{ij}, the resulting effective matrix $\widehat{W}_\mathrm{a}$ would have a spectral radius approximately equal to one. Our goal was, however, to find a dynamic scaling rule for a_i that relied solely on local intrinsic quantities of the corresponding neuron.

A first step towards such a rule is to note that the circular law can be extended to the case where the rows or columns of the matrix do not have uniform variances. Rajan and Abbott [124] considered square matrices with columns having different means and variances. If excitatory weights balance along the rows of the matrix, the square of the spectral radius is $\langle \sigma^2_{\mathrm{w},i} \rangle_i$, where $\sigma^2_{\mathrm{w},i}$ are the variances of the columns. The eigenvalues of the matrix do not change under a transposition, therefore the same result applies to matrices with $\sigma^2_{\mathrm{w},i}$ representing row-wise variances and an E-I balance along the columns. This balance is not a strict assumption in our model, but, given the assumptions we made on the underlying distribution of W_{ij}, the average deviations of the row-wise mean from zero scale as $1/\sqrt{N}$, becoming negligible for large N. Therefore, we can estimate the spectral Radius R_a of the effective matrix by

$$R^2_\mathrm{a} \approxeq \left\langle R^2_{\mathrm{a},i} \right\rangle_i \,, \qquad\qquad R^2_{\mathrm{a},i} \equiv a_i^2 \sum_j W_{ij}^2 \tag{3.14}$$

for large N. We refer to $R^2_{\mathrm{a},i}$ as the *local estimates* of the squared spectral radius R^2_a. Alternatively, this estimate can be expressed using the Frobenius norm as

$$R^2_\mathrm{a} \approxeq \left\| \widehat{W}_\mathrm{a} \right\|^2_\mathrm{F} / N \,. \tag{3.15}$$

For $N = 500$, we evaluated the estimate in (3.14) for a_i drawn from a uniform distribution on $[0, 1]$ and found an average relative error of approximately 3.5%. Since the quantities $R^2_{\mathrm{a},i}$ are essentially neuron-specific estimates of the spectral radius, we concluded that the correct global spectral radius is attained if the population average over these local estimates matches the target.

If we recall (2.60) as the description of the expected variance of the recurrent membrane potential for the mean field theory, we can state a slightly different local variant by replacing the population average in (2.60) by a temporal average (and including the scaling factors a_i):

$$\left\langle x^2_{\mathrm{r},i}(t) \right\rangle_t = a_i^2 \sum_{j,k=1}^{N} W_{ij} W_{ik} \left\langle y_j(t) y_k(t) \right\rangle_t . \tag{3.16}$$

At this point, the equation is still exact. Similar to the argument used to derive the population average, we now assume that the neuronal activities are uncorrelated, meaning that $\langle y_j(t) y_k(t) \rangle_t = \delta_{jk} \langle y_j^2(t) \rangle_t$ (and assuming zero mean activity):

$$\left\langle x^2_{\mathrm{r},i}(t) \right\rangle_t = a_i^2 \sum_{j=1}^{N} W_{ij}^2 \left\langle y_j^2(t) \right\rangle_t . \tag{3.17}$$

Given that the synaptic scaling factors a_i have been adjusted according to (3.12) and have reached a stationary configuration, we find as a temporal average

$$\left\langle x_{\mathrm{r},i}^2(t)\right\rangle_t = R_t^2 \left\langle y_i^2(t)\right\rangle_t \ . \tag{3.18}$$

Plugging this result into (3.17), we find

$$R_t^2 \left\langle y_i^2(t)\right\rangle_t = a_i^2 \sum_{j=1}^{N} W_{ij}^2 \left\langle y_j^2(t)\right\rangle_t \ . \tag{3.19}$$

If we furthermore assume that the bare synaptic weights W_{ij} are not correlated with the average square of their corresponding presynaptic activities $\langle y_j^2(t)\rangle_t$, the sum can be written as $\langle y_i^2(t)\rangle_{t,i} \sum_{j=1}^{N} W_{ij}^2$. An additional average over the index i in (3.19) results in

$$R_t^2 \left\langle y_i^2(t)\right\rangle_{t,i} = \left\langle R_{\mathrm{a},i}^2\right\rangle_i \left\langle y_i^2(t)\right\rangle_{t,i} \approxeq R_{\mathrm{a}}^2 \left\langle y_i^2(t)\right\rangle_{t,i} \ , \tag{3.20}$$

showing that $R_{\mathrm{a}}^2 \approxeq R_t^2$ if the adaptation rule has reached a stationary state.

This derivation required two key assumptions: First, neuronal activities should not be correlated. Whether this holds true or not obviously depends on the type of input that the network receives, and we will discuss the effect of interneuronal correlations in Section 3.4.3. A second assumption is that weights are not correlated with the variances of their presynaptic activities. For the random weights considered here, this assumption can be considered true, independent of the input. Especially for sparse recurrent networks, the effect of a weight W_{ij} (that is, projecting from neuron j to neuron i) back onto the activity of the j-th neuron becomes negligible.

3.3 Input Protocols

To determine the effectiveness of the adaptation mechanism as well as for evaluating the task performance after adaptation, we used different types of input protocols. The first distinction concerns the general statistics of the external input. Here, we considered two cases:

- In the first variant, the network was driven by binary input sequences. More precisely, we projected a single binary sequence $u(t) \in \{-1, 1\}$ onto the network using a set of weights $w_{\mathrm{ext},i}$ such that the external input to each neuron was $u(t)w_{\mathrm{ext},i}$. This corresponds to a situation where the network predominantly receives input from e.g. another neuronal ensemble which has an active or inactive state.

- Second, we considered the case where the network receives a superposition of a large number of different independent external sources. Then, the input can be modeled as random independent Gaussian inputs for each neuron. For simplicity, we chose the external input to have zero mean.

In addition to this distinction, we also distinguished between a variant where the variance of the external input was homogeneous across the network, and a situation where the variances were heterogeneous:

- In the homogeneous case, the globally homogeneous input variance was parameterized by the value σ^2_{ext}.
- For the heterogeneous case, a local variance $\sigma^2_{\text{ext},i}$ was randomly drawn from a positive half-normal distribution with the density $p(x) \propto \Theta(x) \exp(-x^2/2\sigma^2_{\text{ext}})$ which led to a population averaged variance given by σ^2_{ext}.

Combined, this gives a total of four different input protocols, "heterogeneous binary", "homogeneous binary", "heterogeneous Gaussian" and "homogeneous Gaussian", that allowed us to study the effect of both cross-correlations in the input as well as heterogeneity in the input strength.

3.4 Results

In Fig. 3.1, we show the dynamics of the spectral radius resulting from the local and global homeostatic adaptation for the four input protocols described in the previous section. The spectral radius was always initialized to $R_{\text{a}} = 2$ and the target was set to $R_{\text{t}} = 1$. For both heterogeneous and homogeneous independent Gaussian input (Fig. 3.1C and Fig. 3.1D), the spectral radius R_{a} was regulated very close to the target. Notably, a high precision for R_{a} is present for heterogeneous input strengths and local adaptation, even if the local estimates $R^2_{\text{a},i}$ of the squared target deviates from this target, as shown in Fig. 3.1D. This is in line with the theory outlined in Section 3.2.1, where we allowed for possible variations in the variances of neuronal activity across the population (see (3.19)), which is a natural consequence of differing external input strengths. The opposite side of this effect can be seen in Fig. 3.1C, where the perfectly homogeneous external input variances lead to very similar local estimates.

In contrast, binary correlated input had a detrimental effect on the correct tuning of the spectral radius if local adaptation was used. This mismatch was completely absent, however, for the global adaptation rule, which can also bee seen in Fig. 3.2, showing the full spectrum of eigenvalues for the final state of the simulation.

To quantify the amount of deviation from the target spectral radius in more detail, we ran a parameter sweep using a range of external input strengths σ_{ext} and three different values for R_{t}, 0.5, 1.0 and 1.5, and calculated the deviation $R_{\text{a}} - R_{\text{t}}$ after adaptation for all four input protocols. The results are shown in Fig. 3.3. The amount of deviation depends approximately linearly on the strength of the external input. Furthermore, smaller target spectral radii amplify the effect. We will return to this observation when discussing the effect of cross-correlated activity on the tuning precision in Section 3.4.3

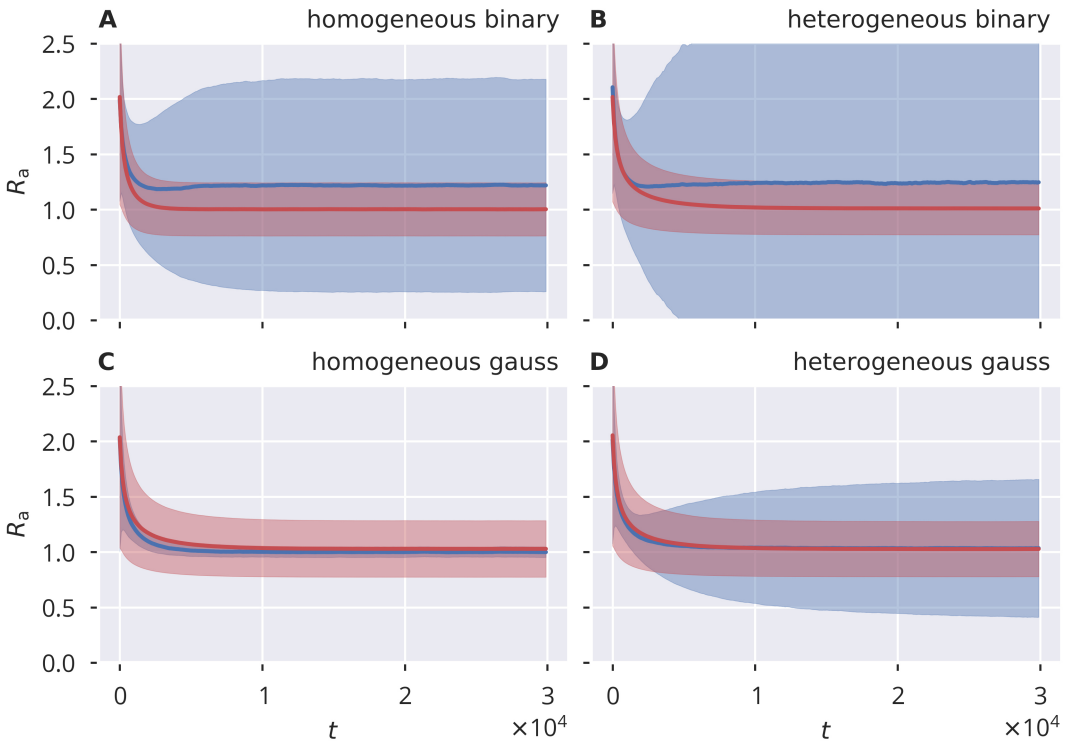

Figure 3.1: Time evolution of the spectral radius R_a and the local estimates $R^2_{\mathrm{a},i}$ for the different input protocols as described in Section 3.3. Panels A–D correspond to homogeneous binary, heterogeneous binary, homogeneous Gaussian and heterogeneous Gaussian input. The solid blue line shows $R_\mathrm{a}(t)$ for local adaptation, whereas the red line shows $R_\mathrm{a}(t)$ for global adaptation. The standard deviation among the local estimates $R^2_{\mathrm{a},i}$ is represented by the widths of the shaded area. The variance of the external input was $\sigma_\mathrm{ext} = 0.5$ for all four simulations, as well as $R_\mathrm{t} = 1$.

3.4.1 Dynamic Mean Field Model and Stabillity

Explaining the theoretical motivation of flow control in Section 3.2.1, we have implicitly assumed that the stationary solution is dynamically stable. In order to better understand whether this holds true, we used a reduced dynamic mean field model that describes the temporal evolution of the population averaged variances of neuronal activity as well as the population average of the synaptic scaling parameters a_i. To do so, we first state that the change of the population average of the the synaptic scaling variable, $a(t) \equiv \langle a_i(t) \rangle_i$, is simply given by the global update rule (3.13) via

$$a(t) = a(t-1)\left[1 + \epsilon_\mathrm{a} \Delta R(t)\right] \ , \qquad \Delta R(t) = R_\mathrm{t}^2 \left\langle y_i^2(t) \right\rangle_i - \left\langle x_{\mathrm{r},i}^2(t) \right\rangle_i \ . \tag{3.21}$$

Using the mean field method outlined in Section 2.3.3 and the approximation (2.68), we arrive at the following dynamic map for the synaptic scaling and population

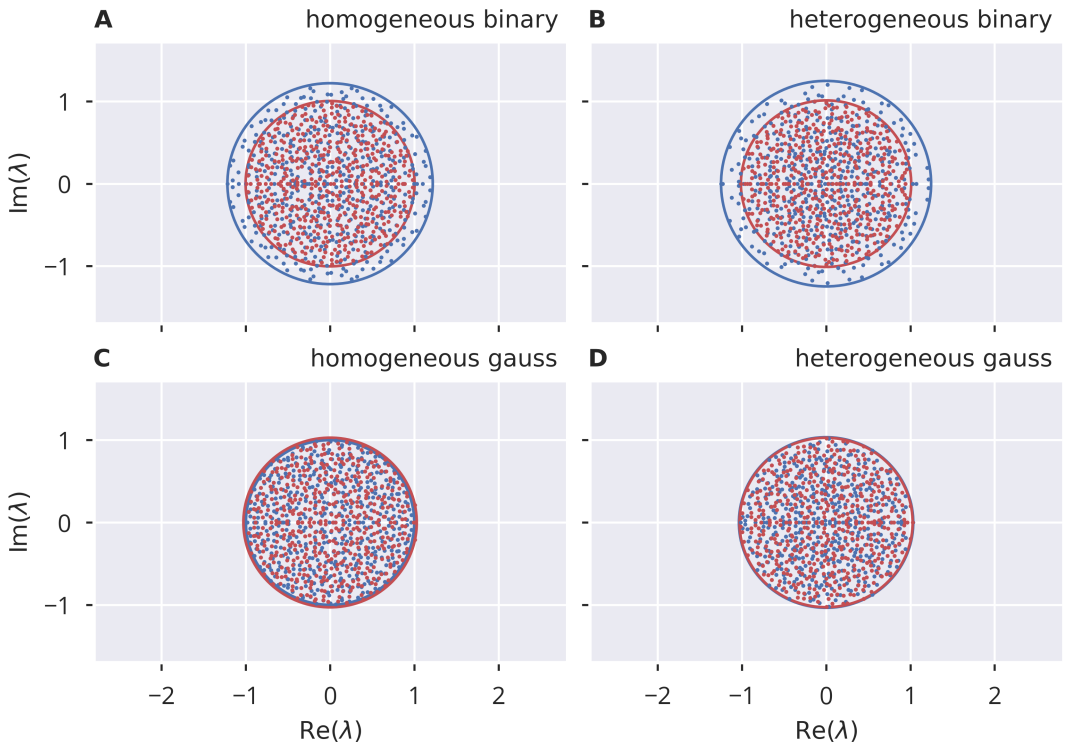

Figure 3.2: Distribution of the eigenvalues of $\widehat{W}_\mathrm{a}$ corresponding to the final state of the simulations shown in Fig. 3.1. Blue: local adaptation. Red: global adaptation.

averaged variance of neuronal activity σ_y^2:

$$a(t) = a(t-1)\left[1 + \epsilon_\mathrm{a}\sigma_\mathrm{y}^2(t-1)\left(R_\mathrm{t}^2 - a^2(t-1)\right)\right] \tag{3.22}$$

$$\sigma_\mathrm{y}^2(t) = 1 - \frac{1}{\sqrt{1 + 2a^2(t-1)\sigma_\mathrm{y}^2(t-1) + 2\sigma_\mathrm{ext}^2(t)}}\,. \tag{3.23}$$

In Fig. 3.4, we have plotted the dynamics of full simulations over the flow of the dynamic mean field model for different parameter values of R_t and σ_ext. Here, we used homogeneous Gaussian input, which makes the external source term $\sigma_\mathrm{ext}^2(t)$ time independent and thus yields an autonomous two-dimensional system. Since the usual choice of the adaptation rate ϵ_a generates much faster dynamics in σ_y^2 than in a, we set $\epsilon_\mathrm{a} = 0.1$ for illustration purposes. Considering that we used an approximation for the calculation of σ_y^2, the dynamic mean field model predicts the transient dynamics and the fixed point (given by the intersection of the nullclines) very well. As a further step, this allowed us to evaluate the stability of the adaptation mechanism on a population level by analyzing the stability of the fixed point of the

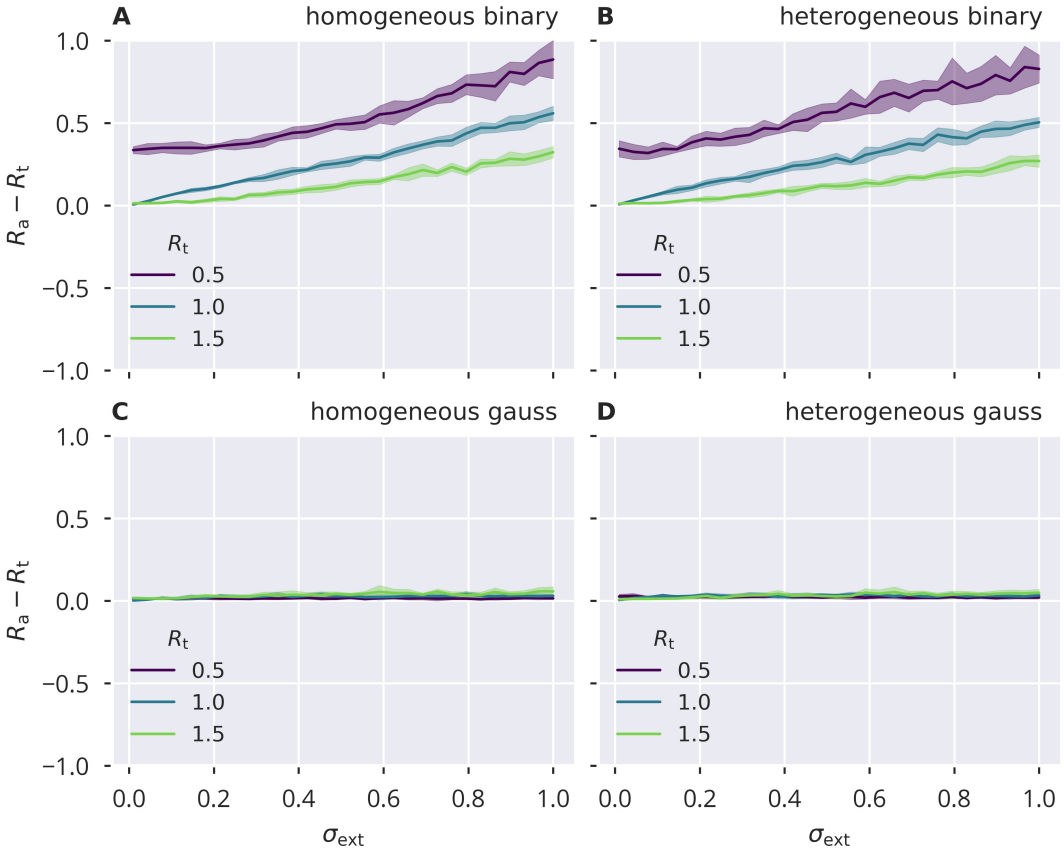

Figure 3.3: Error between the spectral radius R_a and the target R_t after adaptation. Shown are the results for different input strengths, target spectral radii and input protocols. The standard deviation over 10 trials is represented by the height of the respective filled area.

mean field model. The Jacobian matrix of the system is

$$\widehat{J} = \begin{pmatrix} 1 + \epsilon_a \sigma_y^2 \left(R_t^2 - 3a^2\right) & \epsilon_a a \left(R_t^2 - a^2\right) \\ \frac{2a\sigma_y^2}{\left(1+2a^2\sigma_y^2+2\sigma_{ext}^2\right)^{3/2}} & \frac{a^2}{\left(1+2a^2\sigma_y^2+2\sigma_{ext}^2\right)^{3/2}} \end{pmatrix} . \tag{3.24}$$

While the fixed point value for a is simply $a^* = R_t$, it is not possible to find an explicit solution for σ_y^{2*}. Still, it is possible to simplify the Jacobian of the fixed point to

$$\widehat{J}^* = \begin{pmatrix} 1 - 2\epsilon_a \sigma_y^{2*} R_t^2 & 0 \\ 2R_t \sigma_y^2 \left(1 - \sigma_y^{2*}\right)^3 & R_t^2 \left(1 - \sigma_y^{2*}\right)^3 \end{pmatrix} , \tag{3.25}$$

which has eigenvalues $\lambda_1^* = 1 - 2\epsilon_a \sigma_y^{2*} R_t^2$ and $\lambda_2^* = R_t^2 \left(1 - \sigma_y^{2*}\right)^3$. In principle, $|\lambda_1^*|$ can be made greater than one (and thereby making the fixed point unstable) by choosing appropriately large values of R_t, which can, however, be mitigated by a

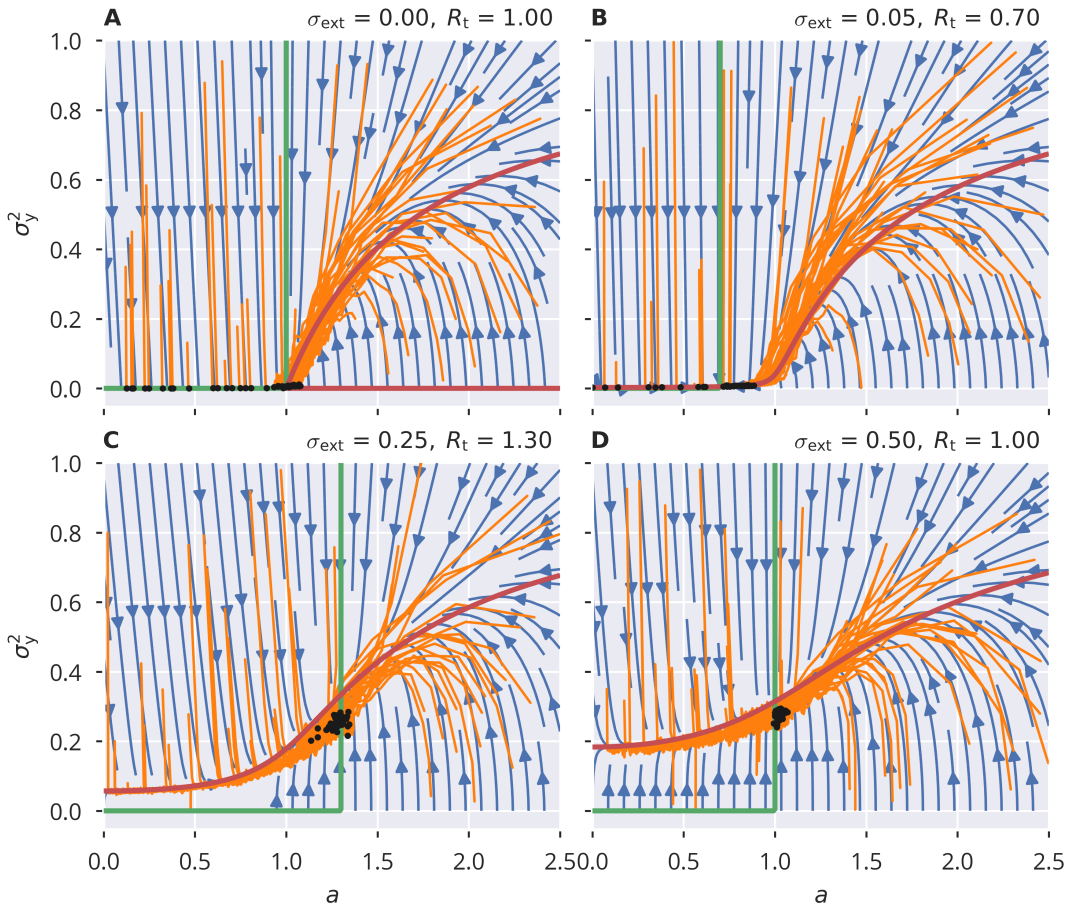

Figure 3.4: Comparison of the dynamic mean field model given by (3.22) and (3.23), represented by the blue flow lines, with full network simulations, shown in orange, for different values of R_t and σ^2_{ext}. Homogeneous Gaussian external input was used. For visualization purposes, we chose a faster adaptation rate of $\epsilon_a = 0.1$. Black dots represent stationary states of the full simulations. The green and red line are the nullclines of the mean field model, i.e. the solutions of $a(t+1) = a(t)$ and $\sigma_y^2(t+1) = \sigma_y^2(t)$, respectively.

slow adaptation rate ϵ_a. For λ_2^*, the picture is less clear, since the behavior depends on the scaling of $\left(1 - \sigma_y^{2*}\right)^3$ relative to R_t^2.

We numerically determined the fixed point of the dynamic mean field model and calculated the eigenvalue spectrum. In Fig. 3.5 we plotted the absolute values of both eigenvalues of $\widehat{J}^*$ for different R_t and σ_{ext} (now using the standard value $\epsilon_a = 10^{-3}$). Both eigenvalues never exceed 1 for the parameter range which we considered realistic and that is shown in the plot. It is interesting to note that λ_2^* approaches 1 for the autonomous network, $\sigma_{ext} = 0$, at the point of the phase transition $R_t = 1$, indicating that, similar to the network dynamics itself, the homeostatic system is also most sensitive to perturbations at this critical point.

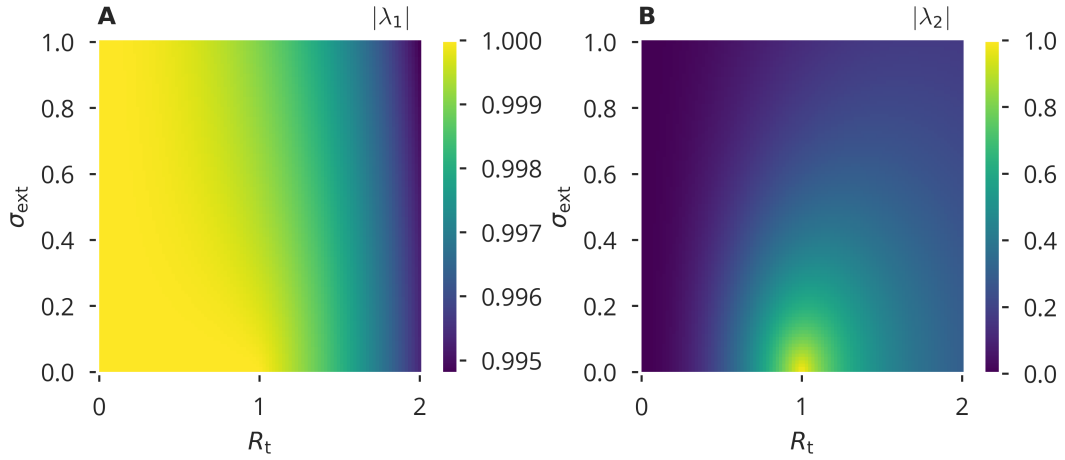

Figure 3.5: Magnitude of eigenvalues of the linearized mean field dynamics given by (3.22) and (3.23) at the fixed point. The adaptation rate of the synaptic scaling was set to the standard value $\epsilon_a = 10^{-3}$.

From this analysis, we concluded that for stationary input statistics, flow control is stable for the parameter range considered to be biologically plausible.

3.4.2 Task Performance

So far, we have only inspected the effectiveness of flow control by itself, without taking its actual effect on network performance into account. To address this point, we constructed a sequential learning task that we termed *XOR-memory recall*. Similar to the pure short-term memory task, see (3.9), the task requires a certain amount of dynamic short term memory, to an extent controlled by the delay parameter τ. However, our goal was to add an additional layer of complexity to the task by including a nonlinear operation. This was done by choosing the random binary sequence described in Section 3.3 for generating the external input and training the network on a target sequence $f_\tau(t)$ given by an XOR operation on subsequent elements of the sequence, delayed by τ:

$$f_\tau(t) = \text{XOR}\left[u(t-\tau), u(t-\tau-1)\right] \ . \quad \tau = 1, 2, ... \tag{3.26}$$

Crucially, the XOR operation is not solvable by a linear classification, thus requiring the use of the inherent nonlinearity of the reservoir. For the training of the readout weights, we used the offline ridge regression procedure as defined in (3.6) and (3.7). We chose a sample size of $T_{\text{sample}} = 10N = 5000$ and a regularization factor $\gamma = 0.01$.

Analogous to the definition of the memory capacity, we defined the total XOR memory capacity, MC_{XOR} as

$$\text{MC}_{\text{XOR}} \equiv \sum_{k=1}^{\infty} \rho^2 \left(f_k(t), y_{\text{out}}\right) \ , \tag{3.27}$$

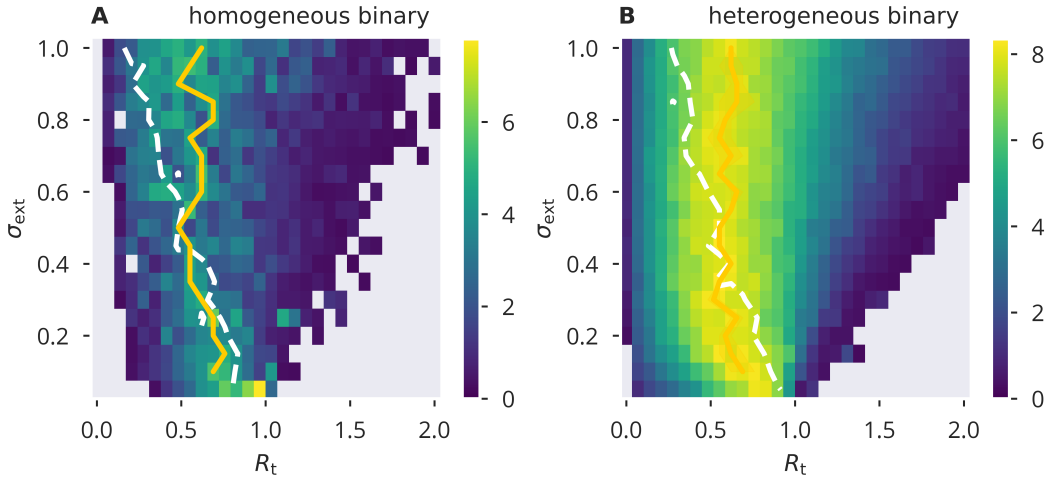

Figure 3.6: XOR memory capacity as defined in (3.27), for networks that were homeostatically adapted using flow control under homogeneous/heterogeneous binary input and different target spectral radii and input variances. The white dashed line denotes pairs of R_{t} and σ_{ext} where $R_{\text{a}} = 1$ after adaptation. The yellow line shows the value of R_{t} for which the performance was maximized for a given σ_{ext}.

again, denoting by $\rho^2\left(f_k(t), y_{\text{out}}\right)$ the squared Pearson correlation between $f_k(t)$ and y_{out}.

Before the activity data was collected for learning, a simulation with homeostatic adaptation was run under the same type of input sequence as in the described task until a stationary state was achieved. Thereafter, the activity batch $\widehat{Y}$ for training the readout weights was collected from a network run without adaptation.

Similar to our analysis of the adaptation mismatch shown in Fig. 3.3, we ran a parameter sweep over σ_{ext} and R_{t} and measured MC_{XOR}. The result is shown in Fig. 3.6. In accordance to the observations shown in Fig. 3.3A/B, the white dashed line denoting $R_{\text{a}} = 1$ deviates from $R_{\text{t}} = 1$ as the input strength increases. Yet, the best performance was achieved for $R_{\text{t}} \approx 0.6$, independent of the external input strength and the actual spectral radius R_{a}. Overall, homogeneous binary input led to a lower performance.

As previously discussed, driving the network exclusively with a single binary sequence represents an "edge case" that is unlikely to mimic input received by local recurrent networks in the cortex. A more realistic scenario could be described by a situation where adaptation takes place under a number of different input streams, while temporarily, a single input source could dominate. To this end, we implemented a variant of the simulation where we adapted the network under independent Gaussian input and then tested the network performance using binary input sequences with the same σ_{ext}. This is shown in Fig. 3.7. Similar to the previous results, optimal performance was achieved for a value of R_{t} that was not significantly affected by σ_{ext}, though being overall larger than the optimal value for binary adaptation.

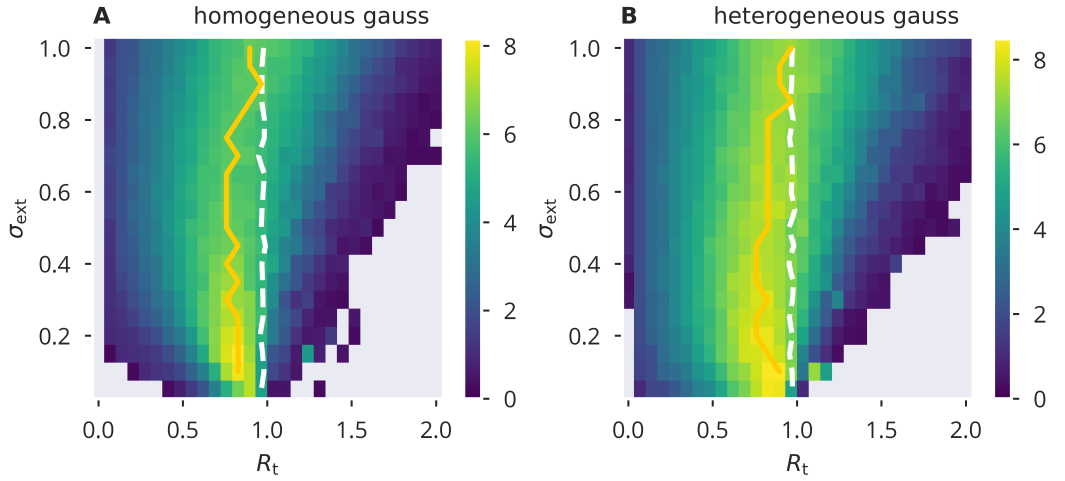

Figure 3.7: XOR memory capacity as given by (3.27). Networks were homeostatically adapted using homogeneous/heterogeneous Gaussian input and then trained and tested using binary input sequences under the same external input variance σ_{ext}. The white dashed line marks $R_{\text{a}} = 1$ and the yellow line shows the value of R_{t} where performance was maximized for a given σ_{ext}.

3.4.3 Cross-Correlations induced by Input

As we have discussed in Section 3.2.1, we can not expect flow control to exactly match the target spectral radius if cross-correlations in the neuronal activity is present, i.e. if $\langle y_i(t)y_j(t)\rangle_t \neq 0$ for $i \neq j$. If we define $x_{\text{bare},i} = \sum_j W_{ij}y_j$ as the "bare" recurrent membrane potential, without the synaptic scaling factors, the essential assumption we made in deriving the adaptation mechanism can be expressed by the statement that the temporal variance of this quantity is given by $\sigma^2_{\text{bare},i} = \sigma^2_{\text{w},i}\sigma^2_{\text{y}}$, where σ^2_{y} is the temporal and population variance of the neuronal activity and

$$\sigma^2_{\text{w},i} \equiv \text{Var}\left[\sum_{j=1}^{N} W_{ij}\right] \tag{3.28}$$

is the variance of the summed bare synaptic weights. Thus, deviations from this prediction of $\sigma^2_{\text{bare},i}$ are expected to negatively affect the precise tuning of the spectral radius. Naturally, even if no correlations are present, we expect this estimate only to become exact for large N due to finite size fluctuations. Therefore, we evaluated the behavior of the population averaged error between $\sigma^2_{\text{bare}} = \langle\sigma^2_{\text{bare},i}\rangle_i$ and $\sigma^2_{\text{w}}\sigma^2\text{y} = \langle\sigma^2_{\text{w},i}\rangle_i\sigma^2\text{y}$ for different network sizes and input protocols to see if different input statistics affected the scaling of this error. This is shown in Fig. 3.8 on a log-log plot. As expected, errors decrease by a power law for both uncorrelated Gaussian input protocols as the network size increases. For correlated binary input on the other hand, the error remains relatively large and shows only weak dependence on the network size.

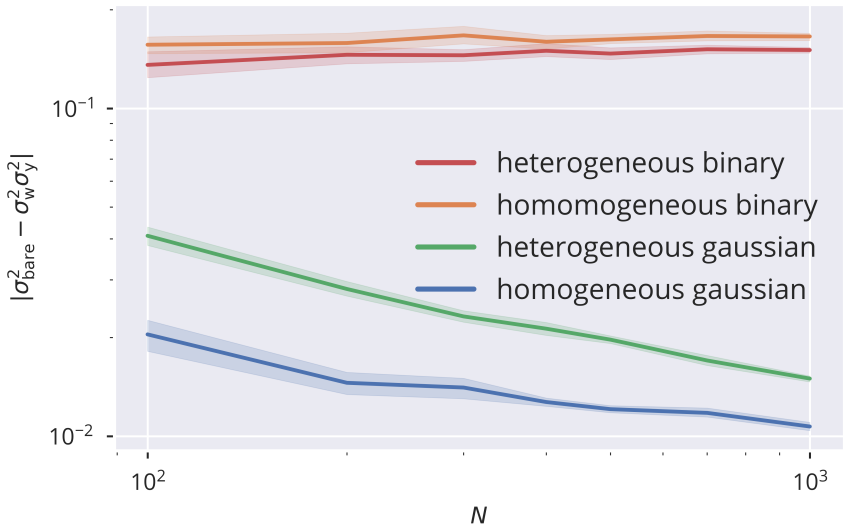

Figure 3.8: Absolute error between population averaged variance of $x_{\text{bare},i} = \sum_j W_{ij} y_j$ and $\sigma^2_{\text{bare}}\sigma^2_{\text{y}}$, see (3.28), for different input protocols, as given in Section 3.3. Convergence can be observed for the input protocols where uncorrelated input is presented to the network. Parameters used in the simulations are $\sigma_{\text{ext}} = 0.5$, $R_{\text{a}} = 1$ and $\mu_{\text{t}} = 0.05$.

To further illustrate that this effect coincides with cross-correlations within the population, we determined the cross-correlation of neuronal activities $\rho(y_i, y_j)$. As an aggregate measure, we defined

$$\overline{\rho^2} \equiv \frac{1}{N(N-1)} \sum_{i \neq j} \rho^2(y_i, y_j) \tag{3.29}$$

as the average squared correlation between the activity of all pairs of neurons in the network. In Fig. 3.9, $\overline{\rho^2}$ is shown as a function of the spectral radius and different external input strengths. A clear distinction can be made between the correlated binary input and uncorrelated Gaussian input. The former causes high cross-correlations, which are more prominent for larger σ_{ext} and get attenuated by increasing the spectral radius, that is, the influence of recurrent coupling. On the other hand, uncorrelated input leads to almost no cross-correlations for small spectral radii, but monotonically increase as a function of R_{a}.

A special case which is also depicted is the case of zero external driving, meaning that any observed cross-correlations are due to autonomous recurrent dynamic. Values for $\overline{\rho^2}$ for $R_{\text{a}} < 1$ should thus be treated with caution, since the actual activity variances are approaching zero even if the network is initialized with non-zero activity. For $R_{\text{a}} > 1$, chaotic dynamics are present in the autonomous case, which corresponds to uncorrelated activity in the limit $N \to \infty$. This implies that the observed cross-correlations can be considered a finite-size effect.

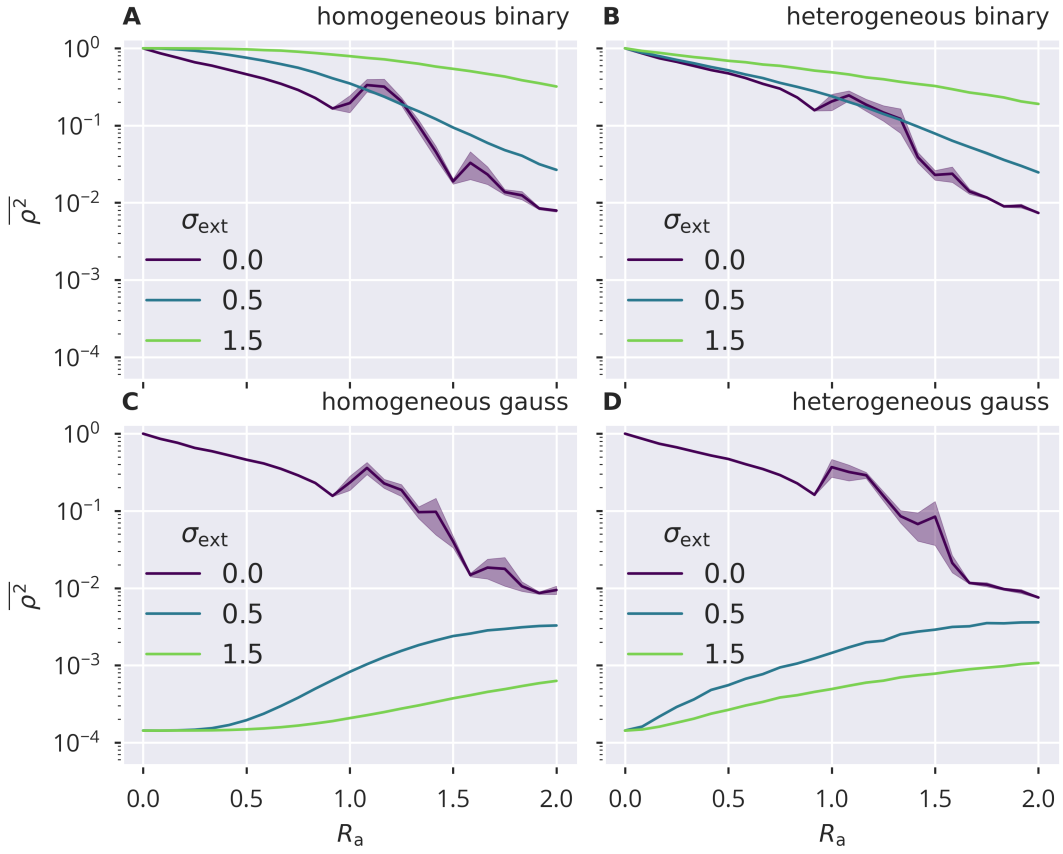

Figure 3.9: Average squared network cross-correlations as defined in (**??**) for different input protocols, spectral radii and input strengths. Averaged over five trials (standard error indicated by shaded area).

3.4.4 Relation between the Tuning Error and Cross-Correlations

Having established that cross-correlations are present under correlated external driving, and that the very same conditions lead to errors in the correct tuning of the spectral radius, it would be desirable to derive a concrete link between those two observations. In appendix A, a detailed analysis of this relation is given, leading to the estimate

$$R_a \approx R_t \sqrt{1 + 2\overline{\rho^2}}\,. \tag{3.30}$$

To test this prediction, we ran the same parameter sweep as used for Fig. 3.3A/B, but also recorded $\overline{\rho^2}$. This allowed us to plot the resulting deviations from the target spectral radius as a function of $\overline{\rho^2}$ instead of the varying σ_{ext} (which, as discussed in the previous section, did still affect $\overline{\rho^2}$). A comparison between the simulation and the prediction by means of (3.30) is shown in Fig. 3.10. The prediction is in excellent agreement if the spectral radius is large enough to decorrelate neuronal activity due to the chaotic recurrent dynamics. For the usual target $R_t = 1$, the simulations using homogeneous input still led to a very good fit to the prediction. For both binary input

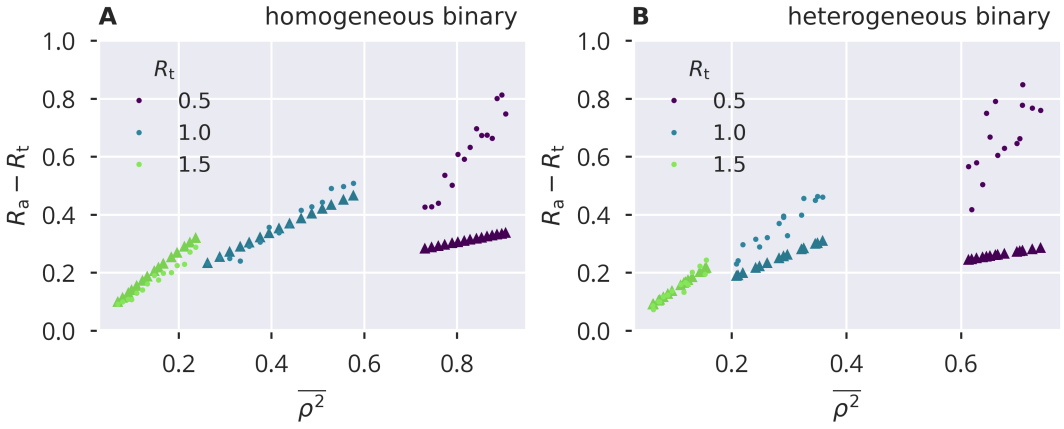

Figure 3.10: Error between the spectral radius of the effective weight matrix $\widehat{W}_a$ and the target spectral radius as a function of the average squared cross-correlation of neuronal activities.

protocols, a small target radius resulted in large cross-correlations, which caused the analytic prediction to substantially deviate.

3.5 Discussion

As indicated by our numerical and theoretical results, the precision of flow control crucially depends on the amount of cross-correlations that are present within the recurrent network. We found the main driver of these correlations to be shared external synaptic input. The extent to which the external driving affected flow control was thus partially determined by the correlations among the external input currents, but also their variance relative to the recurrent input. Thus, the effectiveness of the proposed mechanism depends on the ratio between the fluctuations of recurrent and external input. Binzegger et al. [125] found that approximately 50% of synapses in the rat visual cortex are associated with interlaminar loops and intralaminar connections, which can be considered as being part of a local recurrent structure. For this reason, it is a plausible estimate that the contribution of external and recurrent input is of the same order of magnitude. With respect to flow control, this implies that potential cross-correlations within the external input are to be considered relevant for the tuning of the spectral radius. Furthermore, synchronization can commonly be found in the brain [126] and also might play an important role in processing information [127].

However, correlations in the external input has a detrimental effect on the storage of information in neuronal ensembles [83], and maximal information storage is present if neuronal activity forms an orthogonal ensemble [128, 83, 129]. Moreover, cortical microcircuits exhibit decorrelated firing across neurons in the presence of common external input [130], indicating some form of active cancellation. Hence, the correlations that were found in our network model in the presence of shared external input could potentially be mitigated by changes to the model. For example,

a strict distinction between inhibitory and excitatory neurons could serve the decorrelation of neuronal activity [131, 132]. Furthermore, given that higher dimensional input patterns are presented, plasticity mechanisms could facilitate the emergence of orthogonal representations [128, 83, 133]. Given those potential reductions in cross-correlations, we would thus also expect a more precise tuning of the spectral radius.

Interestingly, despite the observed deviations from the spectral radius, we found a relatively stable network performance over a wide range of external input strengths, see Fig. 3.6. This result might indicate that, while the spectral radius provides a good measure for tuning the properties of neuronal reservoirs, it could be considered a heuristic substitute quantity for the actually more important property of how the neuronal activity is scaled when projected back as recurrent input. Even in the cases where the desired spectral radius was not attained, flow control did, by definition of the stationary solution, enforce the scaling relation $\langle \|\boldsymbol{x}_\mathrm{r}\|^2 \rangle = R_\mathrm{t}^2 \langle \|\boldsymbol{y}\|^2 \rangle$: since the equality $\langle x_{\mathrm{r},i}^2 \rangle = R_\mathrm{t}^2 \langle y_i^2 \rangle$ holds for all i, it is also necessarily true for the sum, which is the squared euclidean norm.

In contrast to conventional homeostatic mechanisms, flow control does not include a pre-defined set point. This loosening of the homeostatic control could be relevant in the face of experimental results, indicating individual homeostatic firing rate set points within a neuronal population [134].

In a broader context, our findings on the behavior of flow control illustrates the potential relevance of the separability of recurrent and external inputs with respect to the understanding of homeostasis in recurrent networks. On the experimental side, homeostasis in neuronal compartments has been investigated to some degree [135, 136, 137, 138], but further theoretical work is needed, especially since there is evidence that the functional segregation across the dendritic structure also has an impact on homeostasis [139].

Apart from introducing a strict distinction between excitatory and inhibitory populations, a natural extension of our model improving biological realism would be the use of spiking neuron models. Firing rate in this respect then becomes an averaged quantity, raising the question of how to properly define higher moments, since longer averaging windows for estimating firing rates also reduces temporal fluctuations of the rate. One "physical correlate" of the firing that is typically studied is the intracellular calcium concentration, since it is, approximately, a temporally filtered trace of the spike train [140]. Cannon and Miller [11] showed that the mean and variance of such a time-averaging physical correlate can be controlled by dual homeostasis, indicating that flow control could also be utilized in spiking networks.

CHAPTER 4

Hierarchical Networks

In the late 1950s, an electrical device was invented by Rosenblatt [141] known as the *perceptron*. In today's terms, it was a single-layered, binary neural network that had the ability to classify input patterns to a certain degree. While its introduction raised the hope of building machines that were able to understand its physical environment on the level of human cognition, in particular recognizing complex visual patterns, it soon became clear that, by mathematical necessity, a single layer of neurons receiving weighted inputs would never be able to correctly classify certain types of input patterns. For example, it was shown that it was incapable of performing a simple XOR operation [142]. In the mid-1980s, the backpropagation was introduced as a breakthrough learning algorithm in the sense that it allowed multiple layers of feed-forward networks, see Fig. 4.1, to be trained on labeled training data [102]. Naturally, the success of this learning approach raised the question as to its applicability in understanding learning in biological networks, and, in its original form, has largely been rejected as a biologically plausible learning algorithm [143, 13, 144].

Despite the criticism regarding the biological plausibility of the learning process, increasing evidence was found that feed-forward network architectures comprised of many hierarchical layers (or *deep networks*, as used in the field of machine learning) are good models for predicting or mimicking patterns of activity that can also be observed in the cortex [145, 146, 147, 148]. Thus, while backpropagation as a learning process is considered biologically implausible, the underlying architecture and the resulting activity patterns are consistent with experimental observations. Consequently, ongoing efforts in theoretical research have been dedicated to finding alternative learning rules that are compatible with biological constraints [144, 149, 105, 150, 151].

One approach to this issue is to consider the unique anatomical properties of cortical pyramidal neurons, and, especially, their dendritic structure. The tree-like structure separating basal, or somatic synapses from the apical, or distal dendritic structure is considered to be suitable for processing functionally different streams of input. This might serve as a physical correspondence to the upstream and downstream of information found in conventional learning algorithms for deep networks [123, 152, 153]. Apart from further understanding the intracellular mechanisms shaping synaptic connectivity as a result of certain different input streams, it also remains an open question what information should actually be sent as "feedback" via up-

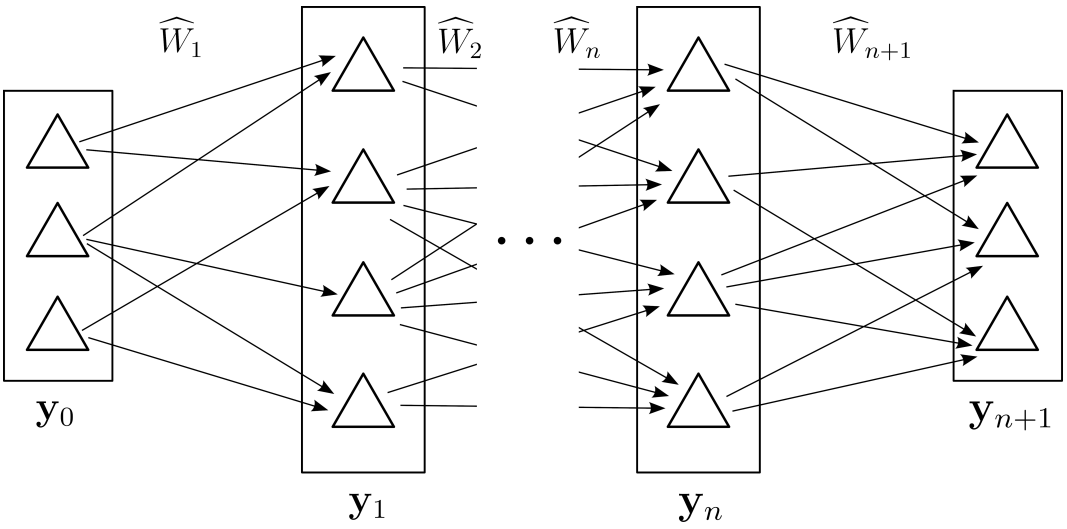

Figure 4.1: Illustration of a multi-layered feed-forward neural network.

stream pathways and how it should be encoded (where by "upstream" in contrast to "downstream" we refer to signals flowing opposite to the usual pathways from earlier stages of sensory processing to areas involved in more complex associations). As a direct analogy to the backpropagation algorithm, some theoretical work was dedicated to biologically plausible algorithms where feedback signals encode learning errors [154, 105, 155, 156]. An alternative approach is known as target propagation: Feedback signals encode targets that are then to be reproduced by the feed-forward pathway [15, 157, 144, 158].

Naturally, depending on the information that is encoded in the feedback signals, different internal plasticity rules are required. In [159], we showed how a simple Hebbian learning rule in combination with a dual homeostatic mechanism in a simple pyramidal compartment model allows feedback signal to serve as target signals for the plasticity in the basal synaptic weights. Before presenting this work in detail, we will first review the anatomical basics that motivates the theoretical work on biologically plausible learning in deep neural networks, which is summarized thereafter.

The mammalian visual cortex is one of the most extensively studied brain region, and, while differences in the processing of other types of sensory input certainly exists by the nature of their spatiotemporal properties, many of the properties and organization of the visual cortex is shared among other cortical regions, such as the auditory cortex [160]. Therefore, we will use the visual cortex as an example case in the following section for introducing the anatomical structure underlying sensory processing.

4.0.1 The Hierarchical Anatomy of the Visual Cortex

In 1962, Hubel and Wiesel [29] performed a series of measurements on the cat visual cortex while presenting a number of basic optical stimuli, such as dots, stripes and bars at different angles. Measuring the activity of single cells in the primary visual cortex, they found a number of cells that reliably responded to stimuli of certain primitive shapes within a confined area on the retina. In contrast, other cells could not be characterized by those simple response properties, and also responded generally to a larger area on the retina. Those experiments were of the first that led to the idea of a hierarchical structure built into cortical visual processing [161]. The receptive fields measured by Hubel and Wiesel are found in the primary visual cortex, V1, which can also be found in the human brain, see Fig. 4.2.

V1 has strong projections onto the adjacent secondary and tertiary visual regions V2 and V3. A popular view on the organization of the visual system is known as the dorsal and ventral stream hypothesis [162]. It describes a split of two streams of information originating from the primary visual cortex, shown in Fig. 4.2 by the upper (dorsal) and lower (ventral) arrows. While the dorsal stream projects from V1 to V3 and further into the parietal lobe (red region), the ventral stream passes through V2 and V4 towards the inferior temporal gyrus (green region). Functionally, those pathways are considered to be responsible for the recognition of objects (ventral) and the identification of their location (dorsal). Within the ventral stream, neurons respond to increasingly complex visual patterns, or features of objects [4, p. 470–472]. While now being considered an over-interpretation of the principle, it was suggested that eventually this functional pathway terminates in single neurons being responsible to entire objects, also termed "grandmother neurons" [163]. A more realistic approach is the idea of objects being represented by populations of neurons corresponding to certain reoccurring features, where neurons with similar responsiveness are physically grouped together [164].

With respect to its connectivity, the dorsal stream is less strictly serial or hierarchical, but, similarly, responses become increasingly complex: In later stages within the parietal lobe, cells respond e.g. to distinct types of motion such as rotation or linear movement [3, p. 334–336].

If one disregards the existence of unique grandmother neurons as being an overly simplistic view in favor of a more parallel approach, the purely feed-forward, hierarchical picture also becomes questionable. In fact, many synaptic connections exist besides those constituting the here presented flow of information. For example, area V3 also has strong connections to V2 and V4 [165].

Functional Hierarchy vs. Anatomical Hierarchy

As pointed out by Hilgetag and Goulas [166], some confusion may stem from the use of the term "hierarchy" in the neurological context, and, related to the concept, the idea of "forward" and "backward" connections. The functional hierarchy and the synaptic connections associated with it are referred to as "topological" hierarchy in [166], since it is an order stemming from the number of synaptic transmission steps

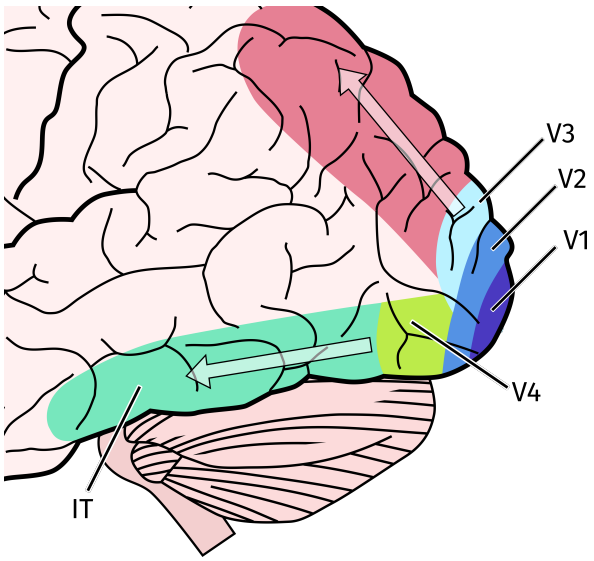

Figure 4.2: Illustration of the visual ventral (lower arrow) and dorsal (upper arrow) stream in the human brain. The ventral stream propagates from the primary visual cortex V1 to V2 and V4, terminating in the inferior temporal gyrus (IT). The dorsal stream propagates from V1 to V3, which then projects into the parietal lobe.

between the source of the sensory stimulus and the respective cortical populations. Alternatively, an often used nomenclature refers to connections as "forward", "backward" and "lateral" based on the origin and destination within the laminar structure of the cortex that as depicted in Fig. 4.3. Connections origination from more superficial layers towards deeper layers are referred to as forward connections, while backward connections are formed in the opposite direction. In addition, intralaminar connections are usually referred to as lateral.

Importantly, the functional hierarchy laid out in the previous section also allows for a distinction between forward connections, projecting from areas of earlier sensory processing to areas with more complex responses, backward, or feedback connections projecting in the opposite direction and, to an extent, lateral connections remaining within a certain cortical area. However, while the functional and laminar perspective on hierarchy may coincide to some degree, there is no necessity that they have to. Still, the assumption that the laminar structure and anatomy of neurons may be involved into the functioning and organization of hierarchical processing remains a valid hypothesis that is considered as a possible explanation for various aspects cognitive functioning including the formation of associations and, more general, the issue of learning and credit assignment [153]. In the following, we will summarize the theoretic approaches to learning and plasticity in hierarchical networks based on the more recent findings on the interplay between the dendritic anatomy of cortical pyramidal neurons and the laminar organization in the cortex. Crucially, these theoretic models indeed assume that, due to a certain regularity in the cortical

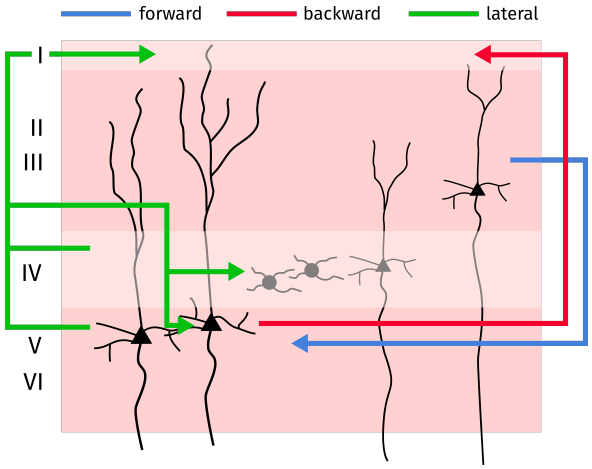

Figure 4.3: Illustration of the laminar structure of the mammalian cortex. Layer I contains mostly dendritic arborizations from deeper cortical areas. Pyramidal neurons are found primarily in Layers IV, V and VI, but also in layer III. Stellate cells can be found in Layer IV. Interlaminar connections can be distinguished into forward, backward, and lateral connections. Source and termination layers are indicated by the colored arrows.

laminar anatomy and connectivity, these anatomical properties can support learning in a functionally hierarchical network.

4.0.2 Models of Biologically Plausible Learning in Hierachical Networks

Before we introduce proposed solutions to the problem of learning and credit assignment in the brain, we shall introduce the mathematical formulation of the classical backpropagation algorithm, which was already briefly mentioned in the beginning of this chapter. Using the nomenclature shown in Fig. 4.1, we define the activity of nodes in a feed-forward network with n intermediate, or "hidden" layers, an input layer $\boldsymbol{y}_0$ and an output layer $\boldsymbol{y}_{n+1}$ as

$$\boldsymbol{x}_i = \widehat{W}_i \boldsymbol{y}_{i-1} \tag{4.1}$$

$$\boldsymbol{y}_i = \phi\left(\boldsymbol{x}_i\right) \tag{4.2}$$

where $1 \leq i \leq n+1$. All layers may have different sizes, which we shall denote by N_i for the ith layer. Given a set $\{(\boldsymbol{y}_0^1, \boldsymbol{f}^1), ..., (\boldsymbol{y}_0^m, \boldsymbol{f}^m)\}$ of m pairs of input vectors $\boldsymbol{y}_0^i$ and target output vectors $\boldsymbol{f}^i$, one can define a loss $\mathcal{L}_i = \mathcal{L}\left(\boldsymbol{y}_{n+1}\left(\boldsymbol{y}_0^i\right), \boldsymbol{f}^i\right)$ whose average over the set of pairs is then to be minimized with respect to all the synaptic weights entering the model:

$$\underset{\widehat{W}_1,...,\widehat{W}_{n+1}}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_i \tag{4.3}$$

The Backpropagation algorithm approaches this minimum by means of gradient descent with respect to each weight. If we use the notation for the element-wise derivative $\boldsymbol{y}'_i \equiv \nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_i)$, we find, for a single pair $(\boldsymbol{y}_0, \boldsymbol{f})$ for $\widehat{W}_{n+1}$

$$\frac{\partial \mathcal{L}}{\partial W_{n+1,ij}} = \frac{\partial \mathcal{L}}{\partial y_{n+1,i}} \left(\boldsymbol{y}_{n+1}\left(\boldsymbol{y}_0\right), \boldsymbol{f}\right) y'_{n+1,i} y_{nj} \ , \tag{4.4}$$

where y_{nj} is the activity of the jth neuron in the nth hidden layer. If we use a squared error loss function, given by

$$\mathcal{L} \equiv \frac{1}{2}\sum_{k=1}^{N_{\text{out}}} e_k^2 = \frac{1}{2}\sum_{k=1}^{N_{n+1}} \left[y_{n+1,k}\left(\boldsymbol{y}_0\right) - f_k\right]^2 \ , \tag{4.5}$$

we can write (4.4) as

$$\frac{\partial \mathcal{L}}{\partial W_{n+1,ij}} = y_{nj} d_{n+1,i} \tag{4.6}$$

$$d_{n+1,i} \equiv y'_{n+1,i} e_i \ . \tag{4.7}$$

Similarly, for all other weights, we get

$$\frac{\partial \mathcal{L}}{\partial W_{l,ij}} = y_{l-1,j} d_{li} \tag{4.8}$$

$$d_{li} \equiv y'_{li} \sum_{k=1}^{N_{l+1}} W_{l+1,ki} d_{l+1,k} \ . \tag{4.9}$$

In vector notation, this can be written as

$$\frac{\partial \mathcal{L}}{\partial \widehat{W}_l} = \boldsymbol{d}_l \boldsymbol{y}_{l-1}^T \tag{4.10}$$

$$\boldsymbol{d}_l \equiv \widehat{D}\left(\boldsymbol{y}'_l\right) \widehat{W}_{l+1}^T \boldsymbol{d}_{l+1} \ , \tag{4.11}$$

with $\widehat{D}\left(\boldsymbol{y}'_l\right)$ denoting a diagonal square matrix with its diagonal elements given by $\boldsymbol{y}'_l$.

While solving (4.1) and (4.2) incrementally from $\boldsymbol{y}_0$ is referred to as the *forward pass*, (4.11) corresponds to an iterative process in the reverse order known as the *backward pass*, or the backpropagation of errors. Depending on the type of activation function, calculating the derivatives entering (4.11) can be done from the stored inputs $\boldsymbol{x}_i$, or directly from the neuronal activities, if a relation such as $\partial/\partial x \tanh(x) = 1 - \tanh^2(x)$ exists. Weights can then updated incrementally by gradient descent with a learning rate ϵ using

$$\Delta W_{l,ij} = -\epsilon \frac{1}{m}\sum_{k=1}^{m} \frac{\partial \mathcal{L}_k}{\partial W_{l,ij}} \ . \tag{4.12}$$

In practice, however, m might be very large, making it impractical to calculate the derivatives for all samples before updating the weights. Faster convergence might be achieved if sub-sampled smaller batches are used to estimate the full gradient. In the extreme case, a single sample is used:

$$\Delta W_{l,ij}(k) = -\epsilon \frac{\partial \mathcal{L}_k}{\partial W_{l,ij}} \ . \tag{4.13}$$

The latter also resembles a more biological approach, since each input pattern and the resulting activity in the network should have an immediate, though very small, effect on the synaptic weights.

The essential issue that renders backpropagation biologically implausible is the fact that the backward pass requires an entire set of mirrored weights $\widehat{W}_i^T$ for each layer to correctly propagate errors. Biologically, this would require an entirely parallel network running in the opposite direction, while tracking changes in the feedforward network, and to date, such mirrored subnetworks were not found. Grossberg [143] referred to this issue as the weight transport problem.

Still, motivated by the success of the algorithm, research has been dedicated to potential workarounds to the problem of weight transport. One proposition to solve this issue was based on the fact that axons also transmit information in the opposite direction, from the synapse to the soma of the presynaptic neuron, potentially removing the need for a separate network entirely [167]. Yet, it was shown empirically that this form of information transmission operates much too slow to serve as a backbone for backpropagation [168]. Despite the lack of experimental evidence, some learning mechanisms have been suggested that would implement backpropagation through an actual separate network [169, 170]. Still, this requires quite complex learning rules to maintain the symmetry between both networks.

Random Feedback Weights

In recent years, another promising approach to biologically plausible backpropagation was was found. The key insight was that learning was still possible without the strict condition of using $\widehat{W}^T$ for the feedback [105, 171]. If certain conditions on the feedback are met, randomly generated feedback weights are sufficient.

While a general condition that would allow to analytically determine learning cases that converge under random feedback weights does not exist to this point, it is possible to derive analytical results under certain simplified assumptions (see [105, Suppl. Note 11]) and Appendix B.2: $\boldsymbol{e}^T \widehat{W} \widehat{B} \boldsymbol{e} > 0$, where $\widehat{B}$ is the random feedback matrix, $\widehat{W}^T$ is the exact backpropagation matrix and $\boldsymbol{e}$ is the feedback error. Note that Lillicrap et al. [105] derived this for linear networks, but showed numerically that successful learning was also possible in more complex, nonlinear networks. In this work, each layer in a multi-layered network had its own random feedback matrix, entering (4.11) instead of $\widehat{W}_{l+1}^T$. This meant that, for nonlinear networks, each layer still included the modulatory factors $\boldsymbol{y}_i' \equiv \nabla_{\boldsymbol{x}} \phi(\boldsymbol{x}_i)$ given by the derivatives. Therefore, biologically speaking, while random feedback as presented in

[105] provides a solution to the problem of weight transport, it still implies that two operations should be carried out at each neuron: the usual nonlinear transformation of the input in the forward pass, as well as providing a multiplicative modulation to the random feedback.

Direct feedback alignment seeks to circumvent the latter issue by providing a feedback to each previous layer instead of using a successive iterative backward pass [155, 172]. Similar to random feedback alignment, such models were also successfully trained on complex classification tasks such as the MNIST dataset [155]. Still, follow-up work found that the trainability of random feedback networks significantly suffers if increasingly deep networks are used [173, 174]. In this case, additional methods are required to increase the alignment between forward and backward weights, which, in a sense, diminishes the appeal of random feedback as a biologically plausible learning framework.

4.0.3 Target Propagation

The previously discussed method of random feedback still assumes that the information that is sent back to earlier stages of information processing encodes errors. This entails another issue in the biological context, being the fact that error signals can be both positive or negative. Yet, if we adhere to a rate encoding scheme of neural processing, neuronal activity is a strictly positive quantity. As a potential workaround, error signals could be transmitted relative to some positive baseline, or, depending on the sign, be sent over separate synaptic pathways. While the latter significantly increases the number of required connections, as well as posing the issue of how such a coding scheme should be temporally coordinated, measuring errors relative to some baseline activity seems unlikely from a metabolic perspective: Neurons would have to constantly be active to a certain degree, even in the absence of any errors.

An alternative to sending error signals as feedback information, known as *target propagation* was first proposed by Le Cun [175] and applied to auto-encoder networks by Bengio [15]. Following studies sought to implement target propagation in biological network models [156, 151, 16]. Here, we briefly summarize the mathematical framework of target propagation as introduced by Bengio [15] and Lee et al. [157].

We adhere to the notation for a feed-forward network as given in (4.1) and (4.2) and first note that for each training input $\boldsymbol{y}_0$, the target $\boldsymbol{l}_{n+1}$ for the last layer $\boldsymbol{y}_{n+1}$ is simply given by the target of the training data, $\boldsymbol{f}$. If we define the mapping from the activity of layer $i-1$ to layer i via

$$\boldsymbol{y}_i = \psi\left(\boldsymbol{y}_{i-1}\right) \equiv \phi\left(\widehat{W}_i \boldsymbol{y}_{i-1}\right) \ , \tag{4.14}$$

we note that, given a target $\boldsymbol{l}_i$ for the activity $\boldsymbol{y}_i$, this target can be attained under a given weight matrix $\widehat{W}_i$ if the activity of $\boldsymbol{y}_{i-1} = \psi^{-1}\left(\boldsymbol{l}_i\right)$, i.e. the inverse of the nonlinear feed-forward transformation. Therefore, an exact backpropagation of targets would be given by

$$\boldsymbol{l}_{i-1} = \widehat{W}_i^{-1} \phi^{-1}\left(\boldsymbol{l}_i\right) \ . \tag{4.15}$$

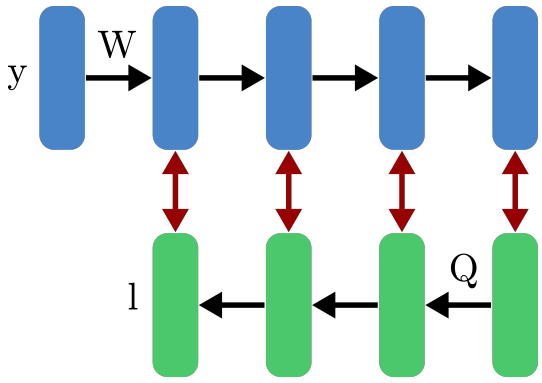

Figure 4.4: Schematic of target propagation. The forward pass of activities $\boldsymbol{y_i}$ is represented by the blue layers, with weights $\widehat{W}_i$. Targets $\boldsymbol{l}_i$ are backpropagated by the feedback weights $\widehat{Q}_i$, allowing for a layer and element-wise loss between targets and activities (red arrows).

In the original target propagation approach, (4.15) is replaced by an approximation such as

$$\boldsymbol{l}_{i-1} = \tilde{\phi}\left(\widehat{Q}_i \boldsymbol{l}_i\right) \ , \tag{4.16}$$

where the $\widehat{Q}_i$ serve as the feedback weights and $\tilde{\phi}$ is a smooth nonlinear activation function (which is not necessarily the same as ϕ). This setup constitutes two learning processes in each layer that can be implemented via a local gradient loss: First, feed-forward weights $\widehat{W}_i$ seek to minimize the local error between the target $\boldsymbol{l}_i$ and the activity $\boldsymbol{y}_i$, $\|\boldsymbol{l}_i - \boldsymbol{y}_i\|^2$. Second, the feedback weights $\widehat{Q}_i$ are adapted to minimize the error between the approximate inverse $\tilde{\phi}(\widehat{Q}_i \boldsymbol{y}_i)$ and the actual activity $\boldsymbol{y}_{i-1}$, $\left\|\tilde{\phi}(\widehat{Q}_i \boldsymbol{y}_i) - \boldsymbol{y}_{i-1}\right\|^2$. In a sense, the latter process uses the activity in the network as training samples for optimizing the projection that is required for the backpropagation of targets. It was suggested by Lee et al. [157] to add noise to the activities in order to optimize the mapping with respect to points that might not be present in the sample space of activities. An illustration of the target propagation principle is shown in Fig. 4.4.

Similar to the random feedback backpropagation approach discussed previously, one might ask if the plasticity within the hidden layers actually reduce the output error using target propagation. It was shown by Lee et al. [157] that this is the case under the condition that the inverse mapping $\tilde{\phi}(\widehat{Q}_i \boldsymbol{y}_i)$ is exactly reproducing $\boldsymbol{y}_{i-1}$.

From a biological perspective, one advantage of target propagation stems from the fact that it resolves the aforementioned issue of propagating signed error signals in backpropagation-inspired frameworks: If the inverse projection of targets is sufficiently precise and strictly positive rates are present in the model, the resulting targets should also always remain positive. Furthermore, more recent studies indicate that cortical feedback indeed appears to predict activity induced by sensory input [176, 177, 178], which bears some resemblance to the inverse mapping $\tilde{\phi}(\widehat{Q}_i \boldsymbol{y}_i)$.

Yet, similar to the random feedback of errors, what still remains is the question of how the parallel forward and backward stream of information illustrated in Fig. 4.4 could be implemented biologically. As both pathways are tightly linked with respect to their plasticity processes, it appears unlikely that an entirely separate network of neurons could be responsible for the backpropagation of targets. In a more general sense, both random feedback and target propagation as candidates for biological learning face the problem of how to combine two linked but functionally different signaling routes into a hierarchical network of neurons.

In recent years, potential solutions to this issue were proposed based on the particular dendritic anatomy of pyramidal neurons in L5 [153, 154, 151, 156], see Fig. 4.3. In contrast to the point-like neuron models introduced in Section 2.2, pyramidal neurons in the cortex possess a quite elaborate, tree-like dendritic structure, that lead to complex dynamical properties [123, 179, 180, 181], which can not be accounted for by point neuron models [182, 183]. One important finding was that the apical dendritic tree of pyramidal neurons can behave as a separate synaptic integration zone [123, 184]. Here, we shall briefly introduce the biophysical properties that were incorporated into theoretical network models.

4.0.4 Nonlinear Dendritic Integration in L5 Pyramidal Neurons

Apart from intralaminar dendritic connections, pyramidal neurons in Layer 5 have a relatively large number of distal dendrites that span into the more superficial supragranular layers 1 and 2. Due to the attenuating effect along the dendrites that transmit postsynaptic potentials to the soma [185, 186], one would expect that these distal synaptic connections have a minor effect on the spike initiation at the soma: In the case of passive dendrites, the voltage along the dendrite as a function of time and space is governed by the cable equation [26], which implies an exponential decay of the signal strength as the postsynaptic potential is transmitted towards the cell body. However, different studies have highlighted the potential importance of these distal dendrites for cognition [187, 188, 189]. Therefore, additional mechanisms appeared to be necessary to facilitate the impact of distal synapses onto somatic spiking activity, and, indeed such mechanisms were subsequently uncovered.

One of these mechanisms exceeding passive dendrite dynamics was found to be the initiation of calcium action potentials (also termed Ca^{2+} *spikes*) in the apical dendritic tree [190, 191, 192]. As described in Section 2.2.4, the spike initiation at the soma was explained in the Hodgkin-Huxley model as a result of the voltage-dependent conductance of the sodium and potassium channels. Similarly, dendritic calcium spikes are initiated by currents induced from voltage-dependent calcium channels in the dendrites. In its most simple form, the dendritic tree may be considered as a separate, point like compartment with active spike initiation dynamics, coupled to the basal somatic component [193]. If a calcium spike is triggered in the dendritic tree, it can, in turn, elicit rapid, bursting somatic spiking activity [192]. This effect can be explained by the relatively long lasting (up to 50 ms) plateau-like potentials generated by the calcium spike, which transiently facilitates spike initiation at the

soma [191, 193]. Moreover, the somatic firing rates caused by dendritic spiking exceeds those that can be achieved if only somatic input is presented [192]. This implies that, rather than being a negligible part of somatic firing, dendritic input can, in principle, become the dominant driver of somatic firing, provided some baseline basal synaptic input is present.

Another process affecting the interplay between the somatic and apical compartment is known as *backpropagation activated* Ca^{2+} *firing* (BAC firing). The previously described spiking mechanism in the dendritic tree can hardly be triggered by solely applying apical synaptic input, but is relatively easy to evoke when somatic spiking is also present [186]. This observation can be explained by somatic action potentials propagating in retrograde direction towards the apical dendritic compartment [183, 123, 153]. Naturally, this bidirectional coupling via both dendritic spiking and BAC firing brings about quite intricate intracellular dynamics. In a more general sense, three major aspects can be identified:

- The apical compartment can be considered an active spike initiation zone, responding to synaptic input nonlinearly.
- Dendritic spiking can transiently depolarize the somatic compartment, facilitating somatic spiking.
- The *maximally* achievable transient firing rates in the soma are also elevated due to dendritic spiking.

From these three features, Shai et al. [181] constructed a phenomenological model predicting the somatic firing rate as a function of the total basal and apical synaptic input and found a very good agreement with simulations using a detailed full compartment model. The model predicts the output frequency y based on the total apical input current I_a and the total basal input current I_b using the following expressions:

$$M\left(I_\mathrm{a}\right) = \alpha_1 + \alpha_2 \sigma\left(\left(I_\mathrm{a} - \alpha_3\right)/\alpha_4\right) \tag{4.17}$$

$$T\left(I_\mathrm{a}\right) = \beta_1 + \beta_2 \sigma\left(\left(I_\mathrm{a} - \beta_3\right)/\beta_4\right) \tag{4.18}$$

$$y\left(I_\mathrm{b}, I_\mathrm{a}\right) = M\left(I_\mathrm{a}\right) \sigma\left(I_\mathrm{b} - T\left(\left(I_\mathrm{a}\right)\right)/\gamma\right) \ , \tag{4.19}$$

where $\sigma(x) = 1/(1+\exp(-x))$ is a standard sigmoidal activation function and the parameters $\alpha_{1/2/3}$, $\beta_{1/2/3}$ and γ were fitted to the full compartment model. The nonlinear dendritic response to synaptic inputs is captured by the sigmoidal functions entering (4.17) and (4.18). While M acts as a modulating factor accounting for the dendritic firing elevating the maximal somatic firing rate, T directly enters the activation function in (4.19) as an input source which corresponds to the transient somatic depolarization.

In Fig. 4.5, we plotted an example of $y\left(I_\mathrm{b}, I_\mathrm{a}\right)$, where we matched the parameters to resemble the shape of the activation function that was found in [181] to best fit the full compartment model. Note, however, that we normalized both the output firing and the input current range to $[0, 1]$. In the original publication, the maximally

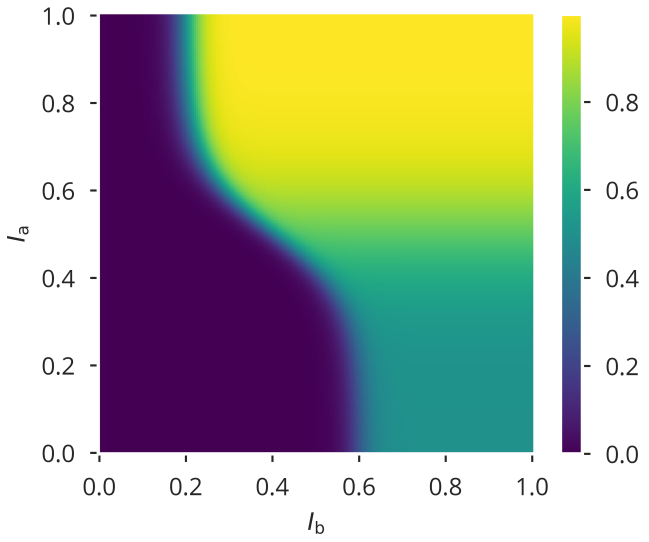

Figure 4.5: Plot of the rate model introduced by Shai et al. [181]. For simplicity, the color-coded output firing rate was normalized to $[0, 1]$. Likewise, the input currents I_b and I_a were rescaled to fit the region of interest into $I_{\mathrm{b/a}} \in [0, 1]$.

achieved firing rate was approximately 150 Hz. Importantly, Fig. 4.5 exhibits two major regions of non-zero activity: an intermediate firing rate that is the result of stimulating only with basal input current I_b, and an area where the maximal firing rate is achieved when stimulating both with I_b and I_a. This distinction between two different modes of firing is considered a potential candidate for detection of temporal coincidence between apical and basal input [153].

Returning to the problem of how errors or targets are backpropagated in a biological network, the insight that the apical dendritic compartment of pyramidal neurons can act as a separate synaptic integration zone inspired theoretical models offering a potential solution to the problem of coordinating coupled forward and backward pathways for learning in hierarchical networks. An important additional insight is the conceptual idea (partially backed up by experimental evidence, see Section 4.0.1) that feed-forward input is projected onto deeper cortical layers, physically close to the basal part of the dendritic structure of layer 5 pyramidal neurons, while feedback signals terminate in more superficial cortical layers, which are physically closer to the apical dendritic tree of the same pyramidal neurons [161, 153].

The combination of the intracellular properties of neurons with segregated integration zones with this topological feature led to a line of research aiming to include those findings into biological models of deep learning. In the following we shall review two such theoretical works.

4.0.5 Deep Learning with Segregated Dendrites

A first model utilizing dendritic segregation in a supervised learning framework was proposed by Sacramento et al. [156]. Here, excitatory pyramidal neurons were modeled as leaky integrators that receive separate inputs from both a somatic and an apical dendritic compartment. In addition, inhibitory interneurons in each layer received input from pyramidal neurons within the same layer, as well as direct input mirroring the activity of neurons in the next layer. Importantly, the apical compartment of the pyramidal neurons receive both top-down excitatory input from the next layer as well as inhibitory input from the interneurons in the same layer. Therefore, the superposition between both inputs encodes an error between the feedback provided by the next layer and the intralaminar prediction via the interneurons. In a pre-training phase, the network is stimulated with random inputs (not providing any specific errors to the last layer) so that the intralaminar inhibitory loop can learn to cancel the feedback provided at the apical compartments. After this stage, the network is considered to be in a "self-predicting state", in the sense that the intralaminar inhibitory pathway predicts the self-generated feedback from each next layer. In a second stage, the activity in the output layer is nudged towards the targets defined by actual training data, and, likewise, the activity of the input layer is also given by the corresponding input pattern. Considering the last hidden layer, the superposition of both top-down input and the input from the inhibitory interneurons in the apical compartments now encode an error between the target activity of the output layer and the activity that would be present in the output layer if it were generated from feed-forward input. This non-zero apical input drives the membrane potential away from the potential induced by the somatic feed-forward input. Learning in this stage now seeks to match the somatic, feed-forward input to the membrane potential nudged by the apical error signal. If this succeeds, the network again has attained a self-predicting state, except that the generated output activity in the last layer now matches the output given by the target activity.

The separation of apical synaptic input in this model serves to calculate an intracellular error signal by means of inhibitory interneurons, while the learning process on the feed-forward can take place separately. Crucially, Sacramento et al. also assume that different learning rules apply to the basal and apical compartment. One advantage of the proposed learning scheme is that the feedback signal simply projects neuronal activities back to the previous layer, while the error information is calculated internally by the superposition with inhibitory input modeling the effect of the feed-forward signal in the next layer.

Another approach to the use of segregated dendrites in hierarchical networks was presented by Guerguiev et al. [151]. As in Sacramento et al. [156], feed-forward inputs are integrated in the somatic compartment, while feedback input arrives at the apical compartment. The resulting membrane potential is then given by a weighted superposition between both inputs, from which a firing rate is calculated. In contrast to Sacramento et al., no inhibitory neurons are used to determine errors. Instead, the learning phase is temporally separated into alternating forward and target phases.

While the forward phase simply calculates the feed-forward and feedback input as well as the activities using only input from the training data, the feedback phase also adds input to the output layer driving the activity towards the desired target output. From these two phases, the authors calculate what they refer to as "plateau potentials", which are nonlinear functions of the average apical input present in each node during the forward and target phase. The intention behind this quantity is to reflect the nonlinear response to apical synaptic input found in pyramidal neurons, as discussed in the previous section. In the spirit of the target propagation approach, the differences between the local plateau potentials in the forward and the target phase define a local error signal that can be used to adjust feed-forward weights. Interestingly, in contrast to the original target propagation approach, it is shown that the method also works with random feedback weights that are not subject to additional plasticity.

Comparing both studies, the presented works solve the coordination of the feed-forward and feedback flow of information in two different ways: While Sacramento et al. introduce an additional inhibitory pathway in each layer, Guerguiev et al. temporally separate the feed-forward and feedback processes. Still, both methods demonstrate the potential advantage of the specific anatomy of pyramidal neurons, making it plausible that synaptic inputs are integrated separately.

Is Learning Driven by Intracellular Error Signals?

What both presented studies have in common is a specific aspect of their learning rules. In both cases, plasticity is driven by the reduction of an internal error signal: In [156], it enters as the difference between the total somatic input and the membrane potential nudged by the apical input. In the study by Guerguiev et al., this error is given by the forward and target plateau potentials. However, a biological plasticity mechanism supporting such error-driven learning is, to our knowledge, not known to date.

Motivated by this issue, we introduced a learning framework attempting to combine the specific dynamics of segregated dendrites in pyramidal neurons with well-documented biological plasticity mechanisms [159]. In the next chapter, we present the results of this study.

CHAPTER 5

Learning by Dendritic Coincidence Detection

Fabian Schubert, Claudius Gros
Nonlinear Dendritic Coincidence Detection for Supervised Learning,
Front. Comput. Neurosci. (2021)
[159]

In the previous chapter, we discussed some of the challenges in transferring machine learning techniques applied to deep networks over to biological neural networks. A relatively new line of research aims to incorporate the distinct physiology of cortical pyramidal neurons. Specific properties of these cells include the physical separation between basal and apical dendrites, the intrinsic neuronal dynamics resulting from this anatomy, and the fact that the overall hierarchical structure in the sensory cortex is, to some degree, reflected in the dendritic termination of feed-forward connections in the basal area and feedback connections in the apical compartment. However, the learning rules devised for these biologically inspired learning frameworks are dependent on an explicit internal error signal and thus, albeit local, different from plasticity mechanisms commonly assumed to be biologically plausible, which usually fall into the class of two-factor Hebbian learning rules [194, 195, 196, 76] or three-factor rules with a modulating component [197, 198, 199, 200]. Hence, we considered the possibility that the nonlinear properties of segregated dendritic integration discussed in the previous section in combination with a conventional Hebbian-type learning rule could yield plasticity dynamics suitable for a supervised learning framework where apical synaptic input provides a feedback signal guiding the plasticity of feed-forward synaptic weights.

Given the potential role of dendritic spiking as a means for coincidence detection, we hypothesized that the strong activity response to temporally coincident basal and distal input could guide plasticity towards an increasing alignment between the basal and apical input. To test this hypothesis, we combined a phenomenological rate model that determines the output firing rate as a function of the total basal and apical input with a standard Hebbian and a BCM-like plasticity rule [201, 202], acting on the basal synapses. If both input streams indeed temporally aligned as a consequence of the plasticity process, the distal input would effectively act as a reconstruction target for the basal inputs. Apart from investigating the dynamics of the plasticity process, we tested our model using a linear supervised binary classification task,

which allowed us to compare the performance to the case of a point neuron subject to the same type of plasticity.

5.1 Model

We studied two types of rate neuron models with two types of plasticity mechanisms, which were tested in all four possible combinations.

5.1.1 Neuron Model

The two-compartment model is a discrete-time rate encoding model including two input variables, the total basal input current I_b and the total apical input current I_a. It is largely based on the phenomenological model of Shai et al. [181] introduced in Section. 4.0.4. To obtain an easily interpretable parameterization, we simplified the model to the following expression:

$$\begin{aligned} y\left(t\right) &= \alpha\sigma\left(I_\mathrm{b}(t) - \theta_\mathrm{b0}\right)\left[1 - \sigma\left(I_\mathrm{a}(t) - \theta_\mathrm{a}\right)\right] \\ &\quad + \sigma\left(I_\mathrm{a}(t) - \theta_\mathrm{a}\right)\sigma\left(I_\mathrm{b}(t) - \theta_\mathrm{b1}\right) \\ \sigma(x) &\equiv \frac{1}{1+\exp(-4x)}\ . \end{aligned} \tag{5.1}$$

The parameters $\theta_\mathrm{b0} > \theta_\mathrm{b1}$ and θ_a denote thresholds for the sigmoidal functions entering the equation. Similar to (4.19), the model consists of two areas of neuronal activity within the two-dimensional space spanned by $(I_\mathrm{b}, I_\mathrm{a})$. These two regions have different maximal firing rates, which are given by 1 and the parameter α, which is always chosen to be within $[0, 1]$. Note that the sigmoidal function was rescaled by a factor of 4 such that the slope at $x = 0$ is 1. A plot of the activation function illustrating the meaning of the parameters is shown in Fig. 5.1.

Similar to the model plotted in Fig. 4.5, two modes of neuronal activity can be distinguished, whose boundaries are defined by the threshold parameters. In the case where both inputs are active, one gets $y \approx 1$. The intermediate activity α is found if only basal input is present. This means that, while maximal activity is obtained if both inputs are active, the neuron can still encode information in the case of basal input alone. It should be noted that, for simplicity, we centered the model around the origin, subtracting potential biases.

We compared the compartment model to a point neuron, which can simply be written as

$$y(t) = \sigma\left(I_\mathrm{b}(t) + I_\mathrm{a} - \theta\right)\ . \tag{5.2}$$

In this study, we only considered a single neuron or layer receiving feed-forward and top-down input. The latter, injected as the apical input, was directly generated to serve as a teaching signal for the learning task at hand. In a multi-layered setup, this input would be a projection of activities from the next network layer. Thus, the

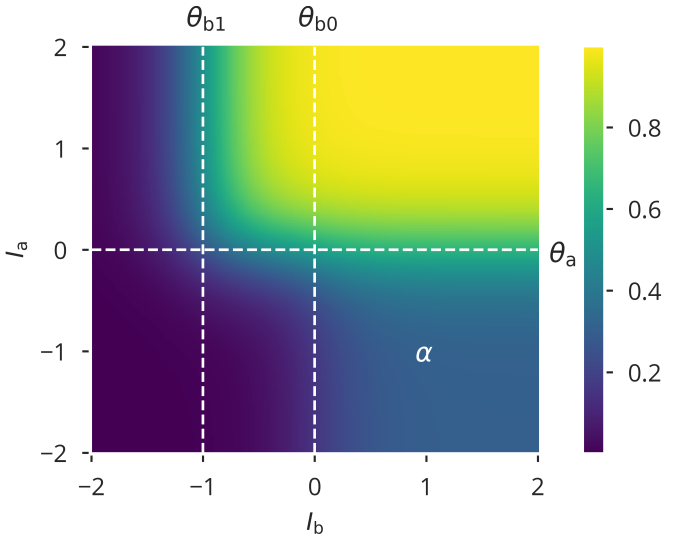

Figure 5.1: Plot of the activation function given by (5.1). The dashed lines indicate the position of the thresholds. The lower right area has a maximal firing rate given by α. Here, $\theta_{b0} = 0$, $\theta_{b1} = -1$,$\theta_a = 0$ and $\alpha = 0.3$.

Table 5.1: Parameter values for the compartment model, homeostatic adaptation and synaptic plasticity.

θ_{b0}	0	σ_a^2	0.25
θ_{b1}	-1	μ_b	0
θ_a	0	μ_a	0
α	0.3	ϵ_b	10^{-3}
ϵ_w	$5 \cdot 10^{-5}$	ϵ_n	10^{-4}
γ	0.1	ϵ_{av}	$5 \cdot 10^{-3}$
σ_b^2	0.25		

apical input current always had the form

$$I_a(t) = n_a(t)x_a(t) - b_a(t) \,, \tag{5.3}$$

with $n_a(t)$ being a scale factor, $x_a(t)$ the teaching signal at hand and $b_a(t)$ a bias. Both the scaling factor and the bias were dynamically adapted according to a dual-homeostatic process, which we describe in Section 5.1.2. Similarly, the basal input $I_b(t)$ was defined as

$$I_b(t) = n_b(t)\boldsymbol{w}^T(t)\boldsymbol{x}_b(t) - b_b(t) \,, \tag{5.4}$$

where $\boldsymbol{w}^T(t)\boldsymbol{x}_b(t)$ denotes the projection of the time dependent, N-dimensional feed-forward input sequence $\boldsymbol{x}_b(t)$ via the basal synaptic weight vector $\boldsymbol{w}(t)$, which is time dependent due to additional synaptic plasticity. As for the apical input, $n_b(t)$ and $b_b(t)$ denote homeostatic scaling and bias parameters. Parameter values can be found in Table 5.1.

5.1.2 Homeostasis

For the biases $b_{\mathrm{b/a}}$ involved in (5.3) and (5.4), the homeostatic adaptation rule reads

$$b_{\mathrm{b/a}}(t+1) = b_{\mathrm{b/a}}(t) + \epsilon_{\mathrm{b}} \left[I_{\mathrm{b/a}}(t) - \mu_{\mathrm{b/a}} \right] \ . \tag{5.5}$$

The inverse adaptation rate $1/\epsilon_{\mathrm{b}} = 10^3$ determines the time scale of the adaptation, driving the inputs towards the target averages $\mu_{\mathrm{b/a}}$. This is similar to the bias adaptation introduced in (3.10) for flow control, except that the bias was coupled to the firing rate rather than the local input current or membrane potential. The variables $n_{\mathrm{b/a}}(t)$ take the role of synaptic scaling factors and are updated according to

$$n_{\mathrm{b/a}}(t+1) = n_{\mathrm{b/a}}(t) + \epsilon_{\mathrm{n}} \left[\sigma^2_{\mathrm{b/a}} - \left(I_{\mathrm{b/a}}(t) - \tilde{I}_{\mathrm{b/a}}(t) \right)^2 \right] \tag{5.6}$$

$$\tilde{I}_{\mathrm{b/a}}(t+1) = (1 - \epsilon_{\mathrm{av}})\, \tilde{I}_{\mathrm{b/a}}(t) + \epsilon_{\mathrm{av}} I_{\mathrm{b/a}}(t) \ . \tag{5.7}$$

The running averages $\tilde{I}_{\mathrm{b/a}}(t)$ are subtracted from the input currents $I_{\mathrm{b/a}}(t)$ in (5.6) to get an estimate of the variances of input currents which are thus dynamically driven towards the targets $\sigma^2_{\mathrm{b/a}}$. Thus, (5.5) and (5.6) constitute two dual homeostatic control loops separately controlling the mean and variance of basal and apical input currents. We chose homeostatic target parameters such that the distribution of apical and basal input serves a certain working regime of the neuronal activation function. For the compartment model considered here, it was set to values that would cover the nonlinearity with respect to both input streams, while preventing saturation, i.e. the case where the activity almost exclusively takes the values 0, α or 1.

Note that for the point neuron model, the same dual homeostatic adaptation was applied for I_{b} and I_{a}. Since the nonlinear sigmoid used in this model was the same as for the compartment model, target parameters remained the same.

5.1.3 Synaptic Plasticity

We considered two plasticity rules that were applied to the basal synaptic weights. The first rule is a simple Hebbian learning rule given by

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \epsilon_{\mathrm{w}} \left[\left(\boldsymbol{x}_{\mathrm{b}}(t) - \tilde{\boldsymbol{x}}_{\mathrm{b}}(t) \right) \left(y(t) - \tilde{y}(t) \right) - \gamma \boldsymbol{w}(t) \right] \tag{5.8}$$

$$\tilde{\boldsymbol{x}}_{\mathrm{b}}(t+1) = (1 - \epsilon_{\mathrm{av}})\, \tilde{\boldsymbol{x}}_{\mathrm{b}}(t) + \epsilon_{\mathrm{av}} \boldsymbol{x}(t) \tag{5.9}$$

$$\tilde{y}(t+1) = (1 - \epsilon_{\mathrm{av}})\, \tilde{y}(t) + \epsilon_{\mathrm{av}} y(t) \ . \tag{5.10}$$

As the trailing averages $\tilde{\boldsymbol{x}}_{\mathrm{b}}$ and $\tilde{y}$ of the basal presynaptic activities and the postsynaptic activity enters (5.8), the change of the basal synaptic weight is determined by the covariance term

$$C(\boldsymbol{x}_{\mathrm{b}}, y) = \left\langle \left(\boldsymbol{x}_{\mathrm{b}}(t) - \langle \boldsymbol{x}_{\mathrm{b}}(t) \rangle_t \right) \left(y(t) - \langle y(t) \rangle_t \right) \right\rangle_t \tag{5.11}$$

and a proportional decay $-\gamma \boldsymbol{w}(t)$, which prevents unbounded growth of the weights. The former covariance term is similar to the Hebbian rule proposed by Linsker [203] in that it is proportional to the correlation between pre- and postsynaptic activities by subtracting the means. Note that, as for the homeostatic plasticity, the averaging time scale is $1/\epsilon_{\mathrm{av}} = 200$. Since our model operates in discrete time steps, we did not explicitly associate a physical time span to each simulation step. However, the transient bursts found by Shai et al. [181] lasted approximately 50–100 ms, which allows us to state that each simulation step corresponds to 0.1 s at maximum. This means that the averaging time scale of 200 steps corresponds roughly to 20 s. Since the temporal adaptation of the average can be considered a form of metaplasticity, which can be rather observed to happen within days [204], it appears that the chosen timescale is too fast. However, this was done for reasons of computational efficiency: increasing the averaging windows does not introduce any new dynamic effects if the time scale is already sufficiently larger than the observed fluctuations around the mean, but requires a longer run-up in the simulation to ensure that the averaging estimate has reached a stationary state. As for the homeostatic model, the Hebbian learning rule could readily applied to both neuron models without changes to the parameters.

BCM Learning Rule

The second plasticity rule that we tested is known as the BCM rule [201, 202]. The key difference to the standard Hebbian learning rule is that, even if pre- and postsynaptic activities are positively correlated, the BCM-rule can induce both long-term potentiation or depression, depending on whether the postsynaptic activity exceeds a threshold or not. The mathematical formulation of the rule applied in our model is given by

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \epsilon_{\mathrm{w}} \left[y(t) \left(y(t) - \theta_{\mathrm{M}} \right) \boldsymbol{x}(t) - \gamma \boldsymbol{w}(t) \right] \ . \tag{5.12}$$

As for the Hebbian rule, the decay term $-\gamma \boldsymbol{w}(t)$ prevents runaway growth of weights. The parameter θ_{M} defines the threshold between long-term depression for $y(t) < \theta_{\mathrm{M}}$ and long-term potentiation for $y(t) > \theta_{\mathrm{M}}$. According to the version of the learning rule proposed by Law and Cooper [205], the threshold is given by $\theta_{\mathrm{M}} = \langle y^2(t) \rangle_t$. Here, the quadratic dependence on the postsynaptic activity was necessary to prevent weights from growing indefinitely. However, since we introduced an additional weight decay, this choice of setting the threshold was not strictly necessary in our model. Importantly, for the compartment model, we explicitly set the threshold value to $\theta_{\mathrm{M}} = (1 + \alpha)/2$. The reasoning behind this choice was to separate the dynamics in the low-activity regime, given by a maximal output rate of α, from the high-activity, bursting regime where the maximal firing rate is 1. By means of this distinction, basal weight configurations that increased the chance of inducing the high-activity regime would be dynamically favored by the plasticity mechanism.

For the point model, this argumentation could not be applied. Yet, for the sake of comparability, we chose to set the threshold θ_{M} to a running average of $y^2(t)$ in this case.

5.2 Results

Depending on the learning task at hand, we used different combinations of basal and apical input patterns. Two protocols were devised. The first, having the purpose of unveiling general properties of the proposed compartment model in combination with the synaptic plasticity rules, is described in Section 5.2.1. In the second protocol, see Section 5.2.2, we sought to further quantify performance differences between the point neuron and the compartment model by means of a binary classification task.

5.2.1 Alignment Between Apical and Basal Inputs

The first simulation protocol was set up to test the ability of both neuron models to align the basal input to some apical teaching input. As a quantifiable measure, we determined the Pearson correlation $\rho(I_{\mathrm{b}}, I_{\mathrm{a}})$ between both input currents after the synaptic plasticity and homeostasis have evolved into a stationary state. For this test, the basal input sequence $\boldsymbol{x}_{\mathrm{b}}(t)$ was independently drawn from a uniform distribution within $[0, 1]$ for each of the N elements and each time step.

For this test to be meaningful, the apical input current should, in principle, be reproducible by the basal input stream. Thus, we constructed the apical input sequence $x_{\mathrm{a}}(t)$ as a linear combination of the basal input sequence via

$$x_{\mathrm{a}}(t) = \boldsymbol{a}^T \boldsymbol{x}_{\mathrm{b}}(t) \tag{5.13}$$

after the basal input sequence was generated, where the projection vector $\boldsymbol{a} \in \mathbb{R}^N$ was generated as a random vector with uniform directional probability and unit length.

Considering the Hebbian learning rule defined in (5.8), we expected that the orientation and dominance of the principal component in the basal input space should significantly influence the resulting stationary weight configuration. Yet, the vector $\boldsymbol{a}$ corresponds to the basal weight configuration resulting in a perfect reconstruction of the apical signal. Consequently, if the direction of the principal component is, in the worst case, orthogonal to $\boldsymbol{a}$, we expected this to act as a strong distraction for the correct reconstruction of the teaching signal, negatively affecting $\rho(I_{\mathrm{b}}, I_{\mathrm{a}})$.

To investigate the effect of distracting input on the teaching signal reconstruction, we added a linear transformation to the input sequence $\boldsymbol{x}_{\mathrm{b}}(t)$, which was parameterized by the scaling factor s, determining the amount of stretching that should be applied orthogonal to $\boldsymbol{x}_{\mathrm{b}}(t)$, and the dimension N_{dist} of the orthogonal distraction subspace. Obviously, N_{dist} could at maximum be equal to $N-1$. The transformation was set up by generating a random orthogonal basis $\widehat{U} \in \mathbb{R}^{N \times N}$ with the constraint that the first basis vector should be parallel to $\boldsymbol{a}$.

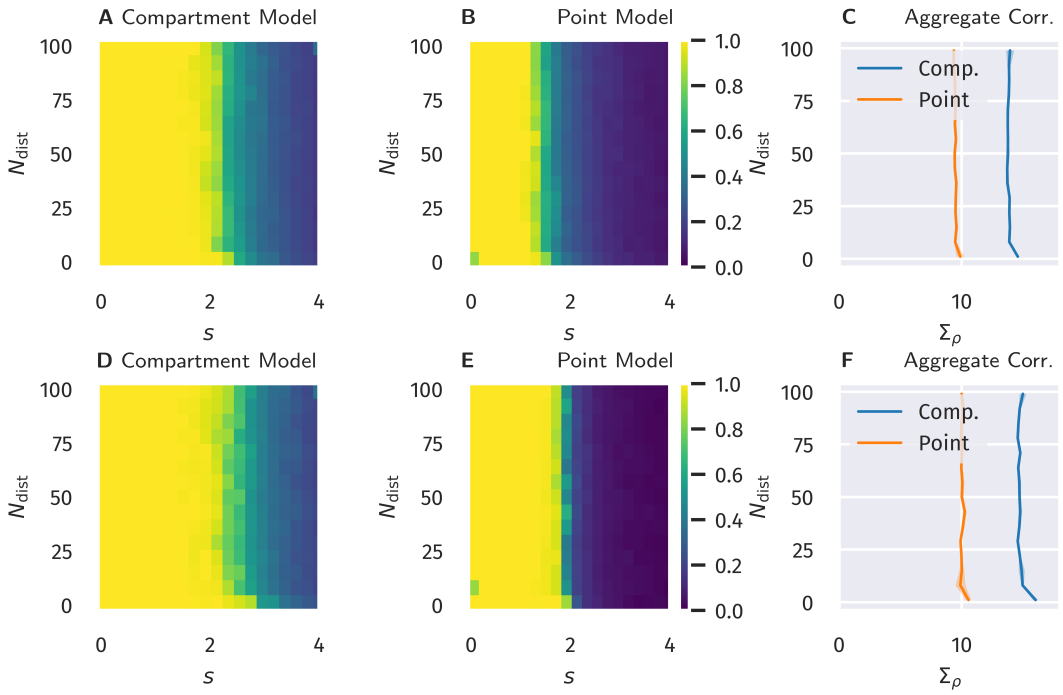

Figure 5.2: Alignment between basal and apical input. The color coding in plots A/B and D/E depicts the Pearson correlation $\rho(I_\text{b}, I_\text{a})$ between basal and apical input after learning. A–C: Using Hebbian learning as given by (5.8). D–F: Using the BCM rule as given by (5.12). Plots C and F depicts the sum of ρ over the s-axis for different N_dist. All depicted values were averaged over 5 trials.

The transformation $\tilde{\boldsymbol{x}}_\text{b}(t) = \widehat{S}\boldsymbol{x}_\text{b}(t)$ was then given by

$$\widehat{S}\boldsymbol{x}_\text{b}(t) = \widehat{U}\,\widehat{D}\,(s, N_\text{dist})\,\widehat{U}^T\boldsymbol{x}_\text{b}(t)\;, \tag{5.14}$$

where $\widehat{D}\,(s, N_\text{dist})$ is a diagonal matrix with $D(s, N_\text{dist})_{ii} = s$ if $2 \leq i \leq N_{2+\text{dist}}$ and $D(s, N_\text{dist})_{ii} = 1$ else. After the learning process has converged using the transformed input $\tilde{\boldsymbol{x}}_\text{b}(t)$, the same input generation procedure was used to determine $\rho(I_\text{b}, I_\text{a})$. During this second phase, plasticity was deactivated. For the simulation, the dimension of the basal input space, N, was set to 100. The same input protocol was also applied when using the BCM-like plasticity rule. In Fig. 5.2, the results are shown for all combinations of neuron models and learning rules.

For all combination of models, we found a relatively quick transition from a fully correlated to a decorrelated configuration as the scaling factor s was increased. However, the dimension of the distraction subspace had only a negligible effect on the resulting alignment. Similarly, only marginal differences could be observed between the two synaptic plasticity rules. Still, the compartment model performed significantly better for both rules, in the sense that it retained the alignment for higher values of s, as illustrated by the aggregate plots shown in Fig. 5.2C/F.

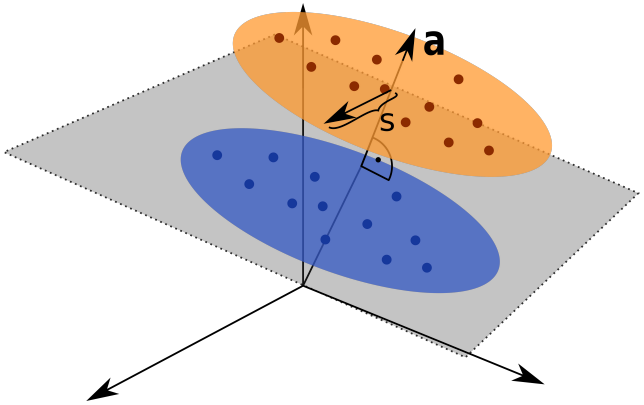

Figure 5.3: Structure of the basal input for the linear classification task. Two Gaussian distributed clusters were generated, located on either side of a separating plane defining the linear classification. The scaling s denotes the standard deviation of the cluster parallel to the plane, with $\boldsymbol{a}$ being its normal vector.

Overall, this first test supported our hypothesis that the particular shape of the compartment model activation function would support the temporal alignment between basal and apical input. Yet, to some degree, we did expect the BCM rule to have a more significant advantage over the Hebbian rule, for the reasons explained in the end of Section 5.1.3.

5.2.2 Performance in a Binary Classification Task

In the second test, we aimed to relate the differences between models described in the previous section to the actual performance in a classification task. To this end, we generated proximal input patterns consisting of two clusters that can be separated by a linear classifier. Fig. 5.3 illustrates the geometry of the basal input. Two Gaussian distributed clusters can be separated by a plane defined by the normal vector $\boldsymbol{a}$, which is a random vector with unit length as introduced for the previous simulation protocol, and an offset vector $\boldsymbol{b}$ whose entries are drawn independently from a uniform distribution on $[0, 1]$. The centers of the Gaussian clusters both have a distance of 0.5 from the separating plane and a standard deviation along the direction of $\boldsymbol{a}$ of $\sigma_{\mathrm{a}} = 0.25$. Orthogonal to $\boldsymbol{a}$, the clusters are Gaussian distributed with standard deviation given by s. In total, each sample can be generated by

$$\boldsymbol{x}_{\mathrm{b}}(t) = \boldsymbol{b} + c(t)\boldsymbol{a} + \widehat{U}\,\widehat{D}\,(s, N_{\mathrm{dist}})\,\widehat{U}^T\boldsymbol{\zeta}(t)\;, \tag{5.15}$$

where $\boldsymbol{\zeta}(t)$ is a random N-dimensional multivariate Gaussian with zero mean and unit variance, $\widehat{U}$ a unitary transform as defined for the previous protocol and used in (5.14). The diagonal matrix $\widehat{D}\,(s, N_{\mathrm{dist}})$ is defined similar to the definition used in (5.14), except that $\widehat{D}\,(s, N_{\mathrm{dist}})_{11} = \sigma_{\mathrm{a}}$, $\widehat{D}\,(s, N_{\mathrm{dist}})_{ii} = s$ for $2 \leq i \leq N_{2+\mathrm{dist}}$ and $\widehat{D}\,(s, N_{\mathrm{dist}})_{ii} = 0$ for $i > N_{\mathrm{dist}}$. Finally, $c(t)$ is a random binary variable with $c(t) \in \{-0.5, 0.5\}$ and equal probabilities, defining whether the sample belongs to

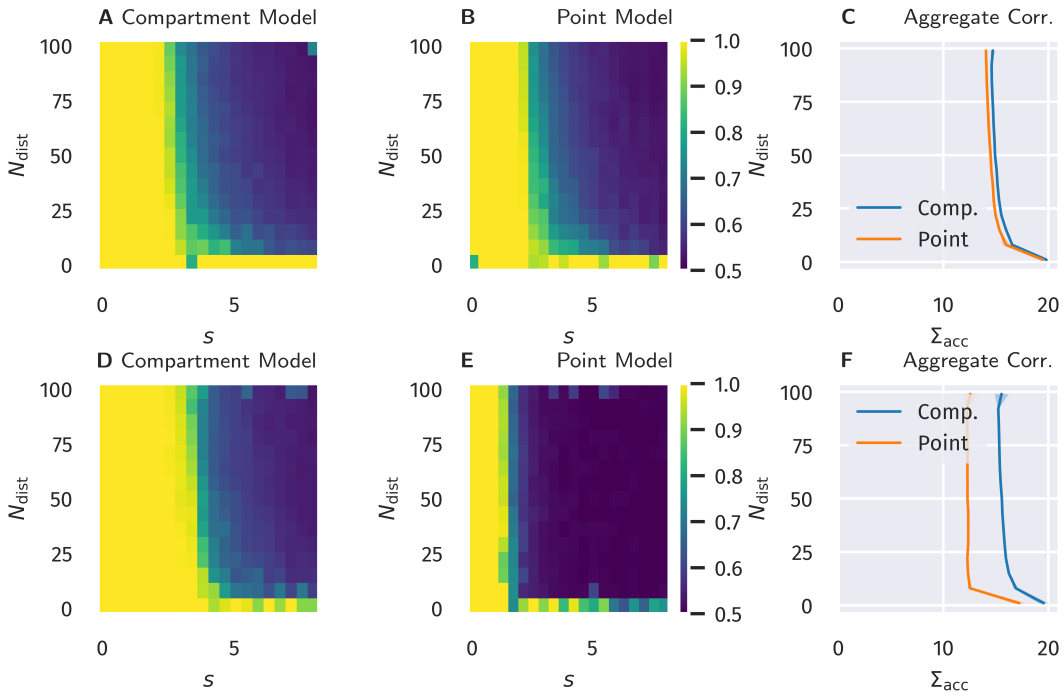

Figure 5.4: Accuracy of binary classification. The color coding depicts the fraction of correctly classified samples for different distraction scaling values s and distraction dimensions N_{dist}. A–C: using Hebbian plasticity. D–F: using the BCM rule. Plot C and F depicts the summed accuracy along the s-axis for different N_{dist} and both neuron models.

either of the two clusters. It should be noted, however, that a small minority of samples fell on the opposite side of the separating plane than the cluster center determined by $c(t)$, but the correct class of each sample was ultimately defined by the location with respect to the separating plane.

For the classification, two neurons were used (numbered with indices 0 and 1), having the task to determine on which side of the separating plane the sample appeared. We used two neurons to allow for a one-hot encoding of this information, which was reflected in the top-down target inputs which were given by

$$x_{\text{a},0}(t) = 1 - \Theta\left(\boldsymbol{a}^T\left(\boldsymbol{x}_{\text{b}} - \boldsymbol{b}\right)\right) \tag{5.16}$$

$$x_{\text{a},1}(t) = \Theta\left(\boldsymbol{a}^T\left(\boldsymbol{x}_{\text{b}} - \boldsymbol{b}\right)\right)\ , \tag{5.17}$$

where Θ denotes the Heaviside step function. After the usual simulation run using plasticity and homeostasis, the classification performance was tested with a simulation run without plasticity/homeostasis and the top-down input turned off. Due to the one-hot encoding, the classification prediction was based on the neuron with the higher activity given the same basal input pattern. The accuracy was defined as the fraction of correctly classified inputs.

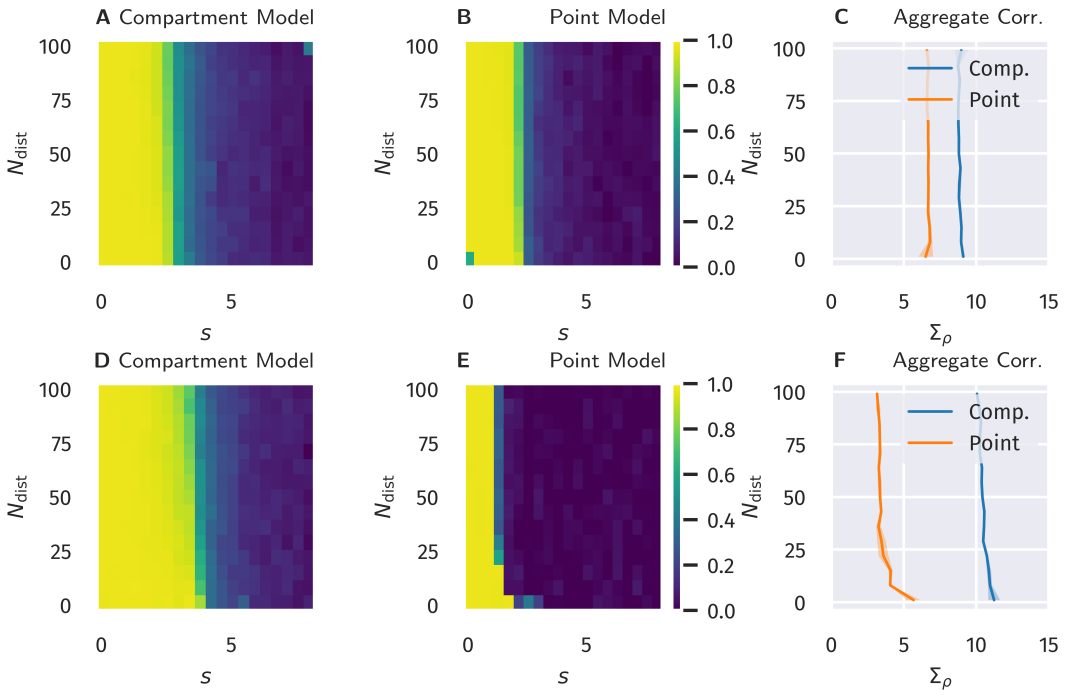

Figure 5.5: Alignment between basal and apical input after training the binary classification task. The plots A/B and D/E show the Pearson correlation $\rho(I_{\text{b}}, I_{\text{a}})$ between basal and apical input after learning. A–C: Using Hebbian learning as given by (5.8). D–F: Using the BCM rule, given by (5.12). Plots C and F depicts the sum of ρ over the s-axis for different N_{dist}. All shown values were averaged over 5 trials.

As shown in Fig. 5.4A–C, Hebbian plasticity did result in a small, yet noticeable difference between the point neuron and the compartment model, where the latter performed better. In contrast, the accuracy using the BCM rule, shown in Fig. 5.4D–F, was significantly improved for the compartment model and also resulted in the overall best performance averaged over the considered parameter space. This result suggested that the compartment model, having two distinct modes of neuronal activation, well suited the particular form of the BCM rule.

Additionally, we quantified the temporal alignment between apical and basal input after learning in the classification by means of the Pearson correlation, as done in the previous setup. The result is shown in Fig. 5.5. Here, the advantage of the compartment model is more pronounced for both plasticity mechanisms. In particular, one can identify parameter values where the point model still showed above-chance classification accuracy while the alignment was very close to zero. We attribute this effect to the method we chose to determine the classification, which was to choose the neuron responding with the higher activity, indifferent of the actual significance of the difference in the activity between both. Thus, as both neurons received the same input patterns, a small difference in the resulting basal input currents might suffice for the correct classification, given that the fluctuations in the

input currents induced by the distracting components of the basal presynaptic input patterns are correlated between both nodes.

5.2.3 Objective Function for BCM Learning

To this point, we have provided some intuitive reasoning regarding the plasticity within the compartment model, in particular for the BCM rule. However, we also wanted to achieve a better theoretical insight into the the learning process driving basal and apical input towards an increased temporal alignment. Therefore, we formulated the BCM rule for the basal weights in the compartment by means of an objective function that is maximized by weight update by stochastic gradient descent.

As a first step, we further simplified the compartment model by replacing the sigmoidal function $\sigma(x)$ with a step function $\Theta(x)$. The overall shape of the model in the space of I_b and I_a is not altered by this replacement, except that the smooth transitions in (5.1) are replaced by sharp edges. If we use the same BCM update rule $\Delta \boldsymbol{w} \propto y(y - \theta_\mathrm{M})\boldsymbol{x}$, we can write the modified update rule as

$$\begin{aligned}\Delta \boldsymbol{w} \propto [&(1-\alpha)\,\Theta\left(I_\mathrm{a} - \theta_\mathrm{a}\right)\Theta\left(I_\mathrm{b} - \theta_\mathrm{b1}\right) \\ &+\alpha\,(\alpha-1)\,\Theta\left(\theta_\mathrm{a} - I_\mathrm{a}\right)\Theta\left(I_\mathrm{b} - \theta_\mathrm{b0}\right)]\,\boldsymbol{x}\ .\end{aligned} \tag{5.18}$$

Conveniently, $\Theta(x)$ is also the first derivative of the ReLu function $[x]^+ \equiv \max(0, x)$. Therefore, we can define the update rule as the first derivative of an objective function $\mathcal{L}_\mathrm{b}$ with respect to the weights:

$$\Delta \boldsymbol{w} \propto \frac{\partial \mathcal{L}_\mathrm{b}}{\partial \boldsymbol{w}} \tag{5.19}$$

$$\begin{aligned}\mathcal{L}_\mathrm{b} \equiv\ &(1-\alpha)\,\Theta\left(I_\mathrm{a} - \theta_\mathrm{a}\right)\left([I_\mathrm{b} - \theta_\mathrm{b1}]^+ + \theta_\mathrm{b1}\right) \\ &+\,\alpha\,(\alpha-1)\,\Theta\left(\theta_\mathrm{a} - I_\mathrm{a}\right)[I_\mathrm{b} - \theta_\mathrm{b0}]^+\ .\end{aligned} \tag{5.20}$$

Note that we are free to include additional terms in $\mathcal{L}_\mathrm{b}$ that are not dependent on $\boldsymbol{w}$, which we did by adding θ_b1 in the first parenthesis, allowing for a better visualization. This objective function is plotted in Fig. 5.6. One can observe that values of $\mathcal{L}_\mathrm{b}$ tend to be higher along the I_b-I_a diagonal. Therefore, if we assume that both input streams have zero mean, we can expect that a high correlation leads to a larger average $\langle \mathcal{L}_\mathrm{b} \rangle_t$.

The relation between the correlation ρ of basal and apical input currents and the expectation value of the objective function can be determined analytically for the case where $\theta_\mathrm{a} = \theta_\mathrm{b0} = 0$, $\theta_\mathrm{b1} \ll -\sigma_\mathrm{b}$ and $(I_\mathrm{b}(t), I_\mathrm{a}(t))$ is modeled as a random vector distributed as a two-dimensional normal distribution with zero mean and a covariance matrix given by

$$\Sigma = \begin{pmatrix} \sigma_\mathrm{b}^2 & \rho\sigma_\mathrm{b}\sigma_\mathrm{a} \\ \rho\sigma_\mathrm{b}\sigma_\mathrm{a} & \sigma_\mathrm{a}^2 \end{pmatrix}\ . \tag{5.21}$$

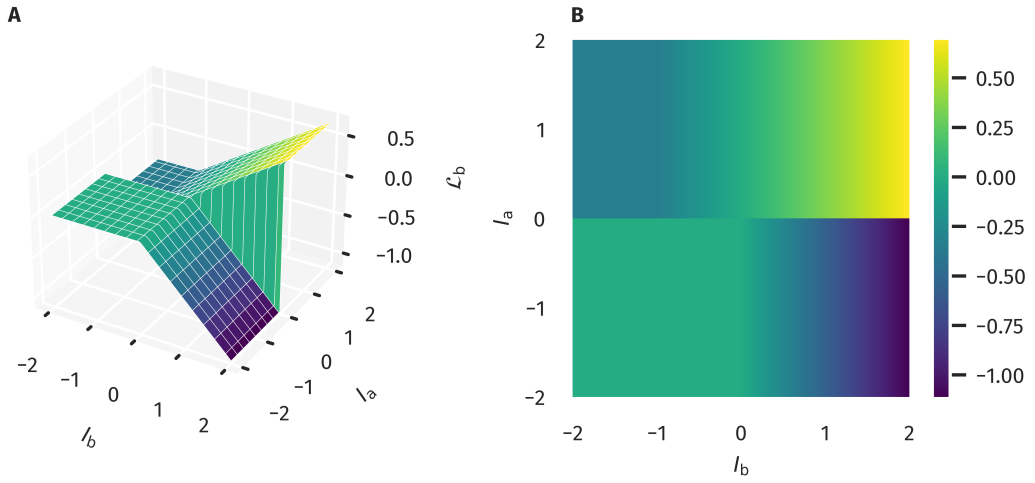

Figure 5.6: Illustration of the objective function $\mathcal{L}_\mathrm{b}$ defined in (5.20) as a 3D plot (A) and color coded (B). The functions tends to be maximized along the I_b-I_a diagonal, thus favoring both inputs being correlated.

In this case, the Gaussian integral can be evaluated, giving

$$\langle \mathcal{L}_\mathrm{b} \rangle_t = \frac{(1-\alpha)\,\sigma_\mathrm{b}}{\sqrt{8\pi}} \left[(2+\alpha)\,\rho - \alpha \right] \ . \tag{5.22}$$

Thus, under these assumptions, the objective function is guaranteed to increase linearly with the basal-apical correlation ρ. Conversely, if the mean and variance of the basal input is kept at their homeostatic targets, an increase of $\langle \mathcal{L}_\mathrm{b} \rangle_t$ must then necessarily go along with a higher correlation. Yet, as we shown in Section 5.2.1, the correlation does not always reach its maximum if distracting input is present. In that respect, it should be noted that the assumptions entering this analysis might not always be sufficiently fulfilled and, in particular, the smoothness of the actual activation function could alter the results under certain inputs. Moreover, even if the temporal average of the objective function monotonically increases, it does not imply that it has to reach the maximum since the rate of increase might approach zero such that $\langle \mathcal{L}_\mathrm{b} \rangle_t$ tends toward a sub-optimal limit from below.

5.2.4 Maximal Correlation vs. Minimal Mean Squared Error

As discussed in Section 4.0.2, many approaches to biological learning via feedback weights implicitly or explicitly strive to reduce a local or global loss function that is usually assumed to be the mean squared error (MSE) between two biologically relatable quantities. In the work presented in this chapter, we argued for a maximization of correlation between basal and apical input I_b and I_a under homeostatic constraints controlling the mean and variance of both variables. Naturally, one might wonder how this approach relates to the minimization of the mean squared error with respect

to the basal weights. Suppose we define a squared error loss function as

$$\mathcal{L}_{\mathrm{SE}} \equiv \left[I_{\mathrm{b}}(t) - I_{\mathrm{a}}(t)\right]^2 \tag{5.23}$$

whose average $\langle \mathcal{L}_{\mathrm{SE}}(t) \rangle_t$ is to be minimized with respect to the basal weights $\boldsymbol{w}$ and an intercept, or bias parameter b. This is, essentially, a linear regression problem [206, chapt. 3] and the question at hand is whether a minimization of $\langle \mathcal{L}_{\mathrm{SE}} \rangle_t$ yields the same result for the weights as a maximization of ρ under certain constraints regarding the mean and variance. For simplicity of the notation, we consider the general case of a linear regression problem where a variable f should be predicted by a linear model $y = \boldsymbol{w}^T \boldsymbol{x} + b$, where $\boldsymbol{x}$ is the input variable, $\boldsymbol{w}$ the linear coefficients and b an intercept parameter. Generally, we find that the relation between $\rho(f, y)$ and $\mathrm{MSE}(f, y) \equiv \langle (f - y)^2 \rangle$ is

$$\rho\left(f, y\right) = \frac{\left\langle f^2 \right\rangle + \left\langle y^2 \right\rangle - 2 \left\langle f \right\rangle \left\langle y \right\rangle - \mathrm{MSE}\left(f, y\right)}{2 \sigma_{\mathrm{f}} \sigma_{\mathrm{y}}} \,, \tag{5.24}$$

where σ_{f} and σ_{y} are the respective standard deviations. First, we note that decreasing the MSE term on the right side appears to increase the correlation. However, this should not be taken for granted, as changes in the model parameters also affect the mean and variance of y which, implicitly, also enter the equation via $\langle y^2 \rangle$, $\langle y \rangle$ and σ_{y}. On the other hand, we can consider the case where the learning procedure attempts to increase the correlation while keeping the first and second moment of the prediction variable y constant by means of some homeostatic mechanism. In this case, the only term on the right side that can change is the MSE. Therefore, increasing the correlation under the described constraints must necessarily decrease $\mathrm{MSE}(f, y)$. Thus, similar to the argument regarding the relation between ρ and $\langle \mathcal{L}_{\mathrm{b}} \rangle$, both optimization processes are monotonically connected, such that, in this case, increasing ρ decreases the MSE, however, this does not imply that the *absolute* optima of both measures are necessarily attained for the same set of model parameters. However, for the linear model considered here, it can be shown that the optimal parameters for the MSE also indeed maximize the correlation. The parameters are given by

$$\boldsymbol{w}_0 = \widehat{C}_{xx}^{-1} \boldsymbol{c}_{fx} \,, \quad b_0 = \langle f \rangle - \boldsymbol{w}_0^T \langle \boldsymbol{x} \rangle \,, \tag{5.25}$$

where $\widehat{C}_{xx}$ is the covariance matrix of the input $\boldsymbol{x}$ and $\boldsymbol{c}_{fx}$ is the covariance vector between the target f and $\boldsymbol{x}$. For both optimization problems, the solution is found by determining the point of zero gradient in the parameters space. The solution for the correlation maximization can be found in Appendix B.3. In fact, while the intercept variable b is required to obtain the solution for the MSE shown in (5.25), the choice of the bias is, by definition, irrelevant for achieving maximal correlation since it only shifts the mean of y. Furthermore, any rescaling of $\boldsymbol{w}_0$ likewise yields maximal correlation, reflecting the fact that the Pearson correlation is also scale invariant with respect to the variances of the random variables.

In our compartment model, the homeostatic mechanisms ensure that the means of both I_a and I_b are equal. Therefore, if learning is successful and the correlation between I_a and I_b is maximized, the resulting basal weight vector must be parallel to the weight vector that one would have achieved by a MSE minimization, though their length might differ. Thus, maximizing the correlation yields the same ratios between synaptic weights as minimizing the MSE.

5.3 Discussion

In this study, we have shown that the nonlinear dendritic integration in the apical dendritic compartment of cortical pyramidal neurons could serve as a driver for basal synaptic plasticity processes maximizing the correlation between top-down and feed-forward input streams. For both considered testing protocols and plasticity mechanisms, the phenomenological compartment model was more robust against distracting patterns in the basal input than a nonlinear point model.

These results are in line with previous studies, suggesting the pyramidal dendritic structure to have a key role in coordinating feed-forward and feedback signaling in hierarchical neural networks [15, 157, 151].

In these previous studies, backpropagation in networks of pyramidal neurons utilized learning rules requiring an explicit error term. Our work indicates that Hebbian-type learning rules in combination with appropriate homeostatic mechanisms can act as a viable and biologically plausible alternative: By learning to maximize the correlation between basal and apical input under homeostatic constraints, the resulting synaptic weight configuration is the same as if the mean squared error was minimized. In a sense, correlation as an objective function is even more flexible, since it allows for arbitrary variances of both input streams, even though we chose to regulate both basal and apical input currents to the same variance.

As the basal input is a linear combination of the basal presynaptic input patterns, maximizing $\rho(I_\mathrm{b}, I_\mathrm{a})$ is a realization of canonical correlation analysis (CCA) [207], even though the apical input space was only one-dimensional in our model since we directly used a scalar input. CCA as a learning objective was investigated by Haga and Fukai [208]. For the model presented by the authors, BCM-like plasticity resulted in an alignment towards the principal component of the input space. CCA was only achieved if a multiplicative term, consisting of the local basal and apical activity, was included. This finding differs from our our results insofar as a multiplicative term was not required to push the basal synaptic weights towards a maximal basal-apical correlation, given that distraction activity was not too dominant. Not requiring a multiplicative term between both compartments is favorable from a biological perspective, as it avoids the necessity to explain an additional coupling between spatially separate loci within the neuron. Ideally, synaptic plasticity rules should only contain locally available information, such as local membrane potentials [209, 210], and pre- and postsynaptic activity, which is the case for both synaptic mechanisms that we implemented.

Beyond single neurons with predefined top-down input, it remains to be tested how the presented framework could operate in a hierarchical network. In particular, it poses the additional question of how to define appropriate top-down target signals and which plasticity rules in the apical compartment would support the formation of those. One potential solution discussed in Section 4.0.2, random feedback weights, could serve as a starting point for further exploration.

CHAPTER 6

General Discussion and Concluding Remarks

In a 2011 review titled "Neuronal homeostasis: time for a change?", O'Leary and Wyllie [211] discuss different facets of current research on neuronal homeostasis, and also bring up the potential effects and implications of homeostatic adaptation taking place as a dendrite- and compartment-specific, distributed process, rather than by means of a centralized control of intracellular properties. The authors speculate that such a control could lead to emergent, potentially more complex dynamical properties of the cell. Furthermore, they note that cell wide, averaged quantities, such as the intrinsic excitability, that have traditionally been the subject of experimental investigations are likely to be the result of the interaction of a multitude of decentralized mechanisms.

In this work, we have presented two models that illustrate the potential role of such a functional segregation of homeostatic processes. In the context of recurrent networks, flow control is an example of a regulatory mechanism that makes use of an intrinsic distinction between recurrent and external inputs, allowing the network to settle into a dynamic state that is required for the optimal processing of time-dependent input. The importance of this separation stems from the homeostatic target that flow control poses onto the amplification of recurrent inputs. It establishes a control loop in the recurrent part of the network that is—to a certain extent—invariant under changes in the external driving.

For hierarchical networks utilizing feedback information, we have proposed a biological learning scheme that also utilizes two separate homeostatic controls, each regulating the mean and variance of the basal and apical input of a cortical pyramidal neuron. Similar to flow control, the functional separation plays an important role here: It allows the neuron to control the moments of both input streams such that their inherent nonlinear interaction affects the synaptic plasticity in a way that lets the apical input act as a target signal for the basal input.

Apart from the intracellular dissociation of homeostatic controls, our results also highlight the potential importance of controlling higher order statistics of intrinsic physical quantities. While theoretical work was done in this respect [83, 107, 92, 9], showing that the computational capabilities of neurons and networks can be affected by adaptation mechanisms that drive neuronal activity towards a desired distribution rather than just a temporal mean, more experimental evidence is needed in this respect: Is it possible to identify biological feedback mechanisms that control e.g. the

temporal variance of some physical quantity in the cell? One issue that might impede the experimental analysis of higher order homeostasis is that it necessarily involves nonlinear control theory, as we have discussed for dual homeostasis in Section 2.4.1: Identifying nonlinear dynamic relationships in experimentally obtained data poses a significantly harder problem compared to a linear model approach [212, cf. chapt. 10]. Furthermore, if temporal recordings are made on a single cell level, the spiking nature of neuronal activity could potentially further complicate the process as it most likely requires some form of temporal filtering of the data [34, p. 12].

The fact that biological neurons generate spikes is also what might be considered a point of critique regarding both models presented in this work: As we have used rate neurons, we did circumnavigate the question as to how the presented adaptation processes could translate to spiking neurons. As mentioned in Section 3.5, firing rates that affect an adaptation process linearly could, in principle, be directly replaced by a physical quantity representing a running average of the spiking activity. A more challenging question concerns the variance or second moment of neuronal activity, as appearing in (3.12) for flow control. An estimate of the second moment using a running average of the spike train will yield different results depending on the time scale of the averaging. Thus, it is likely that an appropriate scaling determined by this time scale would have to be introduced in the adaptation rule to recover the desired dynamics.

Apart from the firing rate appearing in flow control, it also directly utilizes the local (recurrent) membrane potential. Likewise, input currents are homeostatically controlled in the compartment model. For a spiking neuron model, this could be represented by the temporal dynamics of the sub-threshold membrane potential. Of course, a homeostatic feedback that is directly coupled to the membrane potential is inevitably also influenced by the intermittent occurrence of action potentials. Yet, those events take place within approximately 2–5 ms. Thus, for a neuron spiking at an average rate of e.g. 20 Hz, the actual percentage of time spent within the execution of an action potential is only roughly 5–10 %, thus not substantially affecting the homeostatic feedback.

Outlook

Dual homeostasis controlling the mean and variance of separate intrinsic neural variables as an underlying principle has allowed us to postulate new potential mechanisms for the functioning and adaptation of single neurons and networks. As we have shown their effectiveness within the modeling framework presented in this work, what remains to be investigated is whether the proposed ideas can be transferred into more elaborate and biologically realistic models.

For flow control, future work should investigate the principle in recurrent spiking networks using strictly excitatory and inhibitory units. Apart from the previously discussed issue of how to properly define and implement a measure for the variance of neuronal activity in a spiking model, the following questions could be considered:

- In our current version of flow control, excitatory-inhibitory balance was taken for granted. Could this balance also be dynamically achieved by using different synaptic scaling rules for excitatory and inhibitory synaptic weights?
- Can the separation between external and recurrent inputs be justified by using a more detailed dendritic compartment model, assuming that synapses belonging to either of both input streams are spatially organized into different dendritic locations?
- The important role of the spectral radius in optimizing network performance is evident in the echo state framework. How does this translate to spiking networks?

Regarding our proposed learning rules in a nonlinear compartment model, we suggest that the model should be embedded into a hierarchical network. We expect that this could already pose some challenges and scientific questions:

- How are feed-forward and feedback signals temporally coordinated in a multi-layer network? Does an asynchronous forward and backward pass yield the same results as a synchronous, dynamic model?
- How are feedback weights adjusted? As a starting point, is random feedback sufficient for learning?
- How do the rate-based models of synaptic plasticity presented here translate to spike-based models such as spike-timing dependent plasticity?

Naturally, we would expect that a larger network should be able to tackle more complex learning and classification tasks. At some point, it might be possible to combine recurrent, layer-wise network dynamics and adaptation such as flow control with biologically plausible learning rules for feed-forward and feedback projections. Ultimately, sensory input has, in general, a both a temporal and a spatial quality, and it is likely that a complete picture of cortical learning and information processing will only be acquired by building models that account for both of these aspects.

APPENDIX A

Estimation of the Spectral Radius

In Section 3.4.4, we show from simulations that a direct link exists between neuronal activity cross-correlations and the precision of the tuning of the spectral radius. Here, we describe the theory leading to an analytic prediction of the spectral radius resulting from flow control, taking into account cross-correlations in the neuronal activity.

Our goal here is to derive an analytic prediction of the spectral radius estimate

$$R_{\mathrm{a}}^2 \approxeq \left\langle \sum_{j=1}^{N} a_i^2 W_{ij}^2 \right\rangle_i , \tag{A.1}$$

as given in Section 3.2.1. We begin by denoting that a stationary state of flow control fulfills

$$R_{\mathrm{t}}^2 \left\langle y_i^2 \right\rangle_t = a_i^2 \left\langle x_{\mathrm{bare},i}^2 \right\rangle_t \tag{A.2}$$

where $x_{\mathrm{bare},i} = \sum_{j=1}^{N} W_{ij} y_j$. We solve (A.1) for a_i^2:

$$a_i^2 = R_{\mathrm{t}}^2 \frac{\langle y_i^2 \rangle_t}{\langle x_{\mathrm{bare},i}^2 \rangle_t}. \tag{A.3}$$

Therefore, we find that

$$R_{\mathrm{a}}^2 = R_t^2 \left\langle \frac{\langle y_i^2 \rangle_t}{\langle x_{\mathrm{bare},i}^2 \rangle_t} \sum_{j=1}^{N} W_{ij}^2 \right\rangle_i . \tag{A.4}$$

We assume that the variance of W_{ij} scales with $1/N$, which implies that, for large N, the expression converges to

$$R_{\mathrm{a}}^2 = R_t^2 \left\langle \frac{\langle y_i^2 \rangle_t}{\langle x_{\mathrm{bare},i}^2 \rangle_t} \right\rangle_i . \tag{A.5}$$

As a next step, we assume that the network is self-averaging and replace the average over the population in (A.5) by an average over many realizations of the system for

an arbitrary single node:

$$R_{\mathrm{a}}^2 = R_t^2 \left\langle \frac{\langle y_i^2 \rangle_t}{\langle x_{\mathrm{bare},i}^2 \rangle_t} \right\rangle_{\widehat{W},\widehat{W}_{\mathrm{ext}}} . \tag{A.6}$$

To further progress, we have to consider if the average over the product of $\langle y_i^2 \rangle_t$ and $\frac{1}{\langle x_{\mathrm{bare},i}^2 \rangle_t}$ factors out, i.e., if these two factors are statistically independent. To this end, we can write the second moment of the neuronal activity as

$$\langle y_i^2 \rangle_t = \left\langle \phi^2 \left(a_i x_{\mathrm{bare},i}(t) + u(t) W_{\mathrm{ext},i} \right) \right\rangle_t . \tag{A.7}$$

Plugging in the solution of a_i from (A.3), we get:

$$\langle y_i^2 \rangle_t = \left\langle \phi^2 \left(R_{\mathrm{t}} \sqrt{\langle y_i^2 \rangle_t} \frac{x_{\mathrm{bare},i}(t)}{\sqrt{\langle x_{\mathrm{bare},i}^2 \rangle_t}} + u(t) W_{\mathrm{ext},i} \right) \right\rangle_t . \tag{A.8}$$

Despite the fact that this only yields $\langle y_i^2 \rangle_t$ in an implicit way, it shows that $x_{\mathrm{bare},i}(t)$ enters rescaled by the square root of its own second moment. In a large system, $x'_{\mathrm{bare},i}(t) \equiv x_{\mathrm{bare},i}(t) / \sqrt{\langle x_{\mathrm{bare},i}^2 \rangle_t}$ can thus be modeled by a random variable with zero mean and unit variance for all nodes. Therefore, $\langle y_i^2 \rangle_t$ is statistically independent from $\langle x_{\mathrm{bare},i}^2 \rangle_t$ and only determined by the respective values of $W_{\mathrm{ext},i}$. Returning to (A.6), this allows us to state

$$R_{\mathrm{a}}^2 = R_t^2 \left\langle y_i^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}} \left\langle \frac{1}{\langle x_{\mathrm{bare},i}^2 \rangle_t} \right\rangle_{\widehat{W},\widehat{W}_{\mathrm{ext}}} . \tag{A.9}$$

Estimating the second average over the inverse of the squared bare recurrent weights is not straightforward. However, due to Jensen's inequality [213], we have, in general,

$$\left\langle \frac{1}{\langle x_{\mathrm{bare},i}^2 \rangle_t} \right\rangle_{\widehat{W},\widehat{W}_{\mathrm{ext}}} \geq \frac{1}{\left\langle x_{\mathrm{bare},i}^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}}} . \tag{A.10}$$

Note that the expected value $\left\langle x_{\mathrm{bare},i}^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}}$ is given by

$$\left\langle x_{\mathrm{bare},i}^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}} = \left\langle y_i^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}} , \tag{A.11}$$

which allows us to combine (A.9),(A.10) and (A.11):

$$R_{\mathrm{a}}^2 = R_t^2 \left\langle y_i^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}} \left\langle \frac{1}{\langle x_{\mathrm{bare},i}^2 \rangle_t} \right\rangle_{\widehat{W},\widehat{W}_{\mathrm{ext}}} \tag{A.12}$$

$$\geq R_t^2 \left\langle y_i^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}} \frac{1}{\left\langle x_{\mathrm{bare},i}^2 \right\rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}}} \tag{A.13}$$

$$= R_{\mathrm{t}}^2 \,, \tag{A.14}$$

showing at least that R_{a} can only be larger or equal to R_{t}, with equality only if the fluctuations of $\langle x_{\mathrm{bare},i}^2 \rangle_t$ vanish. To progress further, we thus have to find a description of the distribution of $\langle x_{\mathrm{bare},i}^2 \rangle_t$, and from there, estimate the expected value of its reciprocal $1/\langle x_{\mathrm{bare},i}^2 \rangle_t$. The distribution of $\langle x_{\mathrm{bare},i}^2 \rangle_t$ necessarily has a support of $[0, \infty)$, limiting the possible distributions to use as our model. Based on the empirical observation that $\ln(\langle x_{\mathrm{bare},i}^2 \rangle_t)$ resembles a normal distribution, we chose a log-normal distribution as our model. For our particular application, the log-normal distribution has a handy symmetry relation: If x is log-normal distributed, i.e. if $\log(x)$ is normally distributed, it follows that $\log(1/x) = -\log(x)$ is normally distributed with the same variance as $\log(x)$ and a mean of $-\langle \log(x) \rangle$. To calculate the mean of $1/x$, we parameterize the distribution of x by $\mu_{\log}$ and $\sigma_{\log}^2$, denoting the mean and variance of $\log(x)$. The actual mean is then given by

$$\mu_{\mathrm{x}} = \exp\left(\mu_{\log} + \frac{\sigma_{\log}^2}{2}\right) \,, \tag{A.15}$$

which means, by using the symmetry explained above, that the mean of $1/x$ is simply given by

$$\mu_{1/\mathrm{x}} = \exp\left(-\mu_{\log} + \frac{\sigma_{\log}^2}{2}\right) \,. \tag{A.16}$$

Furthermore, the mean and variance of $\log(x)$ can be calculated from the actual mean and variance of x by

$$\mu_{\log} = \log\left(\frac{\mu_{\mathrm{x}}^2}{\sqrt{\mu_{\mathrm{x}}^2 + \sigma_{\mathrm{x}}^2}}\right) = \log\left(\frac{\langle x \rangle^2}{\sqrt{\langle x^2 \rangle}}\right) \tag{A.17}$$

$$\sigma_{\log}^2 = \log\left(1 + \frac{\sigma_{\mathrm{x}}^2}{\mu_{\mathrm{x}}^2}\right) = \log\left(\frac{\langle x^2 \rangle}{\langle x \rangle^2}\right) \,. \tag{A.18}$$

This means that $\mu_{1/\mathrm{x}}$ is simply given by

$$\mu_{1/\mathrm{x}} = \frac{\langle x^2 \rangle}{\langle x \rangle^3} \,. \tag{A.19}$$

We have already stated that $\langle x_{\mathrm{bare},i}^2 \rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}} = \langle y_i^2 \rangle_{t,\widehat{W},\widehat{W}_{\mathrm{ext}}}$, so, we still need to calculate the second moment $\langle \langle x_{\mathrm{bare},i}^2 \rangle_t^2 \rangle_{\widehat{W},\widehat{W}_{\mathrm{ext}}}$. Writing out the weighted summation

of inputs, we get

$$\left\langle \left\langle x^2_{\text{bare},i} \right\rangle_t^2 \right\rangle_{\widehat{W},\widehat{W}_{\text{ext}}} = \left\langle \sum_{j,k,l,m=1}^{N} W_{ij}W_{ik}W_{il}W_{im} \langle y_j y_k \rangle_t \langle y_l y_m \rangle_t \right\rangle_{\widehat{W},\widehat{W}_{\text{ext}}} . \tag{A.20}$$

Strictly speaking, one might find some indirect causal relation between the afferent weights entering the i-th node and the presynaptic activity. However, if the network is sufficiently large and sparse, such potential recurrence effects can be ignored and we can factor out the average over the weights and over the covariance matrices:

$$\left\langle \left\langle x^2_{\text{bare},i} \right\rangle_t^2 \right\rangle_{\widehat{W},\widehat{W}_{\text{ext}}} = \sum_{j,k,l,m=1}^{N} \langle W_{ij}W_{ik}W_{il}W_{im} \rangle_{\widehat{W}} \, \langle \langle y_j y_k \rangle_t \langle y_l y_m \rangle_t \rangle_{\widehat{W},\widehat{W}_{\text{ext}}} . \tag{A.21}$$

For the first term, we find the following:

$$\langle W_{ij}W_{ik}W_{il}W_{im} \rangle_{\widehat{W}} = \begin{cases} 1/N^2 : \begin{cases} j=k=l=m \\ j=k\neq l=m \\ j=l\neq k=m \\ j=m\neq k=l \end{cases} \\ 0 : \text{else} \end{cases} \tag{A.22}$$

For the second term, we find a similar grouping of indices:

$$\langle \langle y_j y_k \rangle_t \langle y_l y_m \rangle_t \rangle_{\widehat{W},\widehat{W}_{\text{ext}}} = \begin{cases} \langle y^2 \rangle^2_{t,\widehat{W},\widehat{W}_{\text{ext}}} : j=k=l=m \\ \left\langle \langle y^2_\alpha \rangle_t \langle y^2_\beta \rangle_t \right\rangle_{\widehat{W},\widehat{W}_{\text{ext}},\alpha\neq\beta} : j=k\neq l=m \\ \left\langle \langle y_\alpha y_\beta \rangle_t^2 \right\rangle_{\widehat{W},\widehat{W}_{\text{ext}},\alpha\neq\beta} : \begin{cases} j=l\neq k=m \\ j=m\neq k=l \end{cases} \\ \langle \langle y_\alpha y_\beta \rangle_t \rangle^2_{\widehat{W},\widehat{W}_{\text{ext}},\alpha\neq\beta} : \text{else} \end{cases} \tag{A.23}$$

Note that we stated earlier that the average square activity of each node is implicitly determined only by the external weights w_i, which are drawn independently. Therefore, the term in the second line, $\langle \langle y^2_\alpha \rangle_t \langle y^2_\beta \rangle_t \rangle_{\widehat{W},\widehat{W}_{\text{ext}},\alpha\neq\beta}$, is actually the same as in the first line, $\langle y^2 \rangle^2_{t,\widehat{W},\widehat{W}_{\text{ext}}}$, since we do not expect any correlations.

Combining (A.21), (A.22) and (A.23), we get

$$\left\langle \left\langle x^2_{\text{bare},i} \right\rangle_t^2 \right\rangle_{\widehat{W},\widehat{W}_{\text{ext}}} = \frac{N(N-1)}{N^2} \left[\left\langle y^2 \right\rangle^2 + 2\overline{c^2_{\text{off}}} \right] \tag{A.24}$$

$$\approx \left\langle y^2 \right\rangle^2 + 2\overline{c^2_{\text{off}}} , \tag{A.25}$$

where in $\langle y^2 \rangle^2 = \langle y^2 \rangle^2_{t,\widehat{W},\widehat{W}_{\text{ext}}}$ we have omitted the averaging notation and

$$\overline{c^2_{\text{off}}} \equiv \left\langle \langle y_\alpha y_\beta \rangle_t^2 \right\rangle_{\widehat{W},\widehat{W}_{\text{ext}},\alpha\neq\beta} \tag{A.26}$$

is the average of the square of the off-diagonal elements of the product matrix $\langle y_\alpha y_\beta \rangle_t$.

Now we can apply (A.19) to $\left\langle x^2_{\mathrm{bare},i} \right\rangle_t$ and get

$$\left\langle \frac{1}{\langle x^2_{\mathrm{bare},i} \rangle_t} \right\rangle_{\widehat{W},\widehat{W}_{\mathrm{ext}}} = \frac{\langle y^2 \rangle^2 + 2\overline{c^2_{\mathrm{off}}}}{\langle y^2 \rangle^3} \,. \tag{A.27}$$

We use this result in (A.9) and get

$$R^2_{\mathrm{a}} = R^2_{\mathrm{t}} \frac{\langle y^2 \rangle^2 + 2\overline{c^2_{\mathrm{off}}}}{\langle y^2 \rangle^2} \,. \tag{A.28}$$

This simple formula that establishes a link between the resulting spectral radius and the presence of non-vanishing correlations between the activity of the nodes. Obviously, this formula also correctly predicts that R_{a} will be exactly R_{t} if no cross-correlations are present. Moreover, it is in agreement with our earlier result that $R_{\mathrm{a}} \geq R_{\mathrm{t}}$. If the variations of the average squared neuronal activity among the population as well as the off-diagonal elements entering c^2_{off} are small and the activities are zero on average, we may further simplify this equation by stating that $\overline{c^2_{\mathrm{off}}}/\langle y^2 \rangle^2 \approx \overline{\rho^2}$, where $\overline{\rho^2}$ was given in Section 3.4.3, resulting in

$$R^2_{\mathrm{a}} \approx R^2_{\mathrm{t}} \left(1 + 2\overline{\rho^2}\right) \,. \tag{A.29}$$

APPENDIX B

Mathematical Derivations

B.1 Solution to the Regularized Least Squares Problem

Here, we give a derivation of the matrix solution given in Section 3.1 for the regularized least squares problem for the linear readout weights.

Using the same matrix notation for the matrices $Y_{ij} = y_j(i)$ and $F_{ij} = f_j(i)$ introduced for (3.5) and (3.7), the minimization problem is given by

$$\arg\min_{\widehat{V}} \left[\left\| \widehat{V}\widehat{Y} - \widehat{F} \right\|_F^2 + \gamma \left\| \widehat{V} \right\|_F^2 \right] , \tag{B.1}$$

with $\|\cdot\|_F^2$ denoting the square of the Frobenius norm. A convenient property of the Frobenius norm is the identity

$$\left\| \widehat{X} \right\|_F^2 = \mathrm{Tr}\left(\widehat{X}^T \widehat{X} \right) \tag{B.2}$$

for real matrices $\widehat{X}$, allowing us to rewrite (B.1) as

$$\arg\min_{\widehat{V}} \left[\mathrm{Tr}\left(\widehat{V}^T \widehat{V} \widehat{Y} \widehat{Y}^T \right) - 2\mathrm{Tr}\left(\widehat{V} \widehat{Y} \widehat{F}^T \right) + \mathrm{Tr}\left(\widehat{F}^T \widehat{F} \right) + \gamma \mathrm{Tr}\left(\widehat{V}^T \widehat{V} \right) \right] . \tag{B.3}$$

For finding the minimal solution, we set the derivative $\partial / \partial V_{ij}$ of each matrix element with respect to the expression in the parenthesis to zero, giving

$$2 \sum_{l,m=1}^{N,T} V_{il} Y_{lm} Y_{mj}^T - 2 \sum_{m=1}^{T} F_{im} Y_{mj}^T + 2\gamma V_{ij} = 0 \, , \forall i, j \, , \tag{B.4}$$

where N denotes the dimension of the neuronal reservoir and T the number of sampled time steps. In matrix notation, this can be rewritten as

$$\widehat{V} \left(\widehat{Y}\widehat{Y}^T + \gamma \widehat{\mathbb{1}} \right) - \widehat{F}\widehat{Y}^T = 0 \, . \tag{B.5}$$

which is solved for $\widehat{V}$ by

$$\widehat{V} = \widehat{F}\widehat{Y}^T \left(\widehat{Y}\widehat{Y}^T + \gamma \widehat{\mathbb{1}} \right)^{-1} , \tag{B.6}$$

concluding the derivation.

B.2 Sufficient Condition for Random Feedback in Linear Networks

In [105, Suppl. Note 11], a condition on the random feedback matrix $\widehat{B}$ in a linear network was given, guaranteeing stability and convergence of the learning process: $\boldsymbol{e}^T\widehat{W}\widehat{B}\boldsymbol{e} > 0$, where $\boldsymbol{e}$ is the error term and $\widehat{W}^T$ is the exact feedback matrix used in the backpropagation algorithm. Here, we briefly summarize the derivation.

Consider a network with one hidden layer, so that the only weights requiring a feedback signal are given by the matrix $\widehat{W}_1$ projecting from the input to the hidden layer. Furthermore, we assume that the network is linear, i.e. $\phi(x) = x$. In the full backpropagation algorithm, the update rule for a single pair of training data would read

$$\Delta\widehat{W}_1 = -\epsilon\widehat{W}_2^T\boldsymbol{e}\boldsymbol{y}_0^T \tag{B.7}$$

and

$$\Delta\widehat{W}_2 = -\epsilon\boldsymbol{e}\boldsymbol{y}_1^T\ . \tag{B.8}$$

The transposed matrix $\widehat{W}_2$ in $\Delta\widehat{W}_1$ is now replaced with a randomly generated matrix $\widehat{B}$. If a solution exists that exactly reproduces the target outputs, it implies that they can be generated from a "target matrix" $\widehat{T}$ via $\boldsymbol{f} = \widehat{T}\boldsymbol{y}_0$. In this case the error term $\boldsymbol{e}$ becomes $\left[\widehat{W}_2\widehat{W}_1 - \widehat{T}\right]\boldsymbol{y}_0$. Therefore, the weight changes over the full set of input samples is given by

$$\left\langle\Delta\widehat{W}_1\right\rangle = -\epsilon\widehat{B}\widehat{A}\widehat{C}_0 \tag{B.9}$$

$$\left\langle\Delta\widehat{W}_2\right\rangle = -\epsilon\widehat{A}\widehat{C}_0\widehat{W}_1^T \tag{B.10}$$

where we have defined $\widehat{A} \equiv \widehat{W}_2\widehat{W}_1 - \widehat{T}$ and $\widehat{C}_0 \equiv \langle\boldsymbol{y}_0\boldsymbol{y}_0^T\rangle$ is the input auto-correlation matrix. Immediately, it follows from (B.9) and (B.10) that $\widehat{W}_2\widehat{W}_1 = \widehat{T}$ is a stationary solution of the learning process, as expected. However, since we have introduced the random feedback matrix $\widehat{B}$, this stationary solution is not necessarily stable in $\widehat{W}_1$. The condition that should be fulfilled is that the loss function averaged over the input samples should monotonically decrease as $\Delta\widehat{W}_1$ is applied:

$$\langle\Delta\mathcal{L}\rangle = \sum_{k,l}\langle\Delta W_{1,kl}\rangle\frac{\partial\langle\mathcal{L}\rangle}{\partial\Delta W_{1,kl}} < 0 \tag{B.11}$$

In vector notation, this equals

$$\mathrm{Tr}\left(\Delta\widehat{W}_1^T\frac{\partial\langle\mathcal{L}\rangle}{\partial\widehat{W}_1}\right) < 0\ . \tag{B.12}$$

if we further define $\widehat{\Lambda} \equiv \widehat{A}\widehat{C}_0$, (B.12) can be written as

$$-\epsilon \mathrm{Tr}\left(\widehat{\Lambda}^T \widehat{B}^T \widehat{W}_2^T \widehat{\Lambda}\right) = -\epsilon \mathrm{Tr}\left(\widehat{W}_2^T \left\langle \boldsymbol{e}\boldsymbol{e}^T \right\rangle \widehat{B}^T\right) < 0 \,. \tag{B.13}$$

This condition is fulfilled if $\langle \boldsymbol{e}^T \widehat{W}_2 \widehat{B} \boldsymbol{e} > 0 \rangle$, meaning that the error feedback given by $\widehat{B}\boldsymbol{e}$ should not differ by more than a right angle from the exact feedback given by $\widehat{W}_2^T \boldsymbol{e}$. This is necessarily true, independent of the distribution of error signals, if $\widehat{W}_2 \widehat{B}$ is positive definite.

B.3 Solution of Correlation Maximization in a Linear Regression Model

In Section 5.2.4, we discussed the relation between minimizing the mean squared error and maximizing correlation in a linear regression model and stated that the solution for the minimal mean squared error also maximizes the correlation, which is what we will derive here. As in Section 5.2.4, we write the linear model as $y = \boldsymbol{w}^T \boldsymbol{x} + b$ and denote the variable that is to be predicted by f. Defining

$$\mathrm{MSE}(y, f) \equiv \left\langle (y - f)^2 \right\rangle \,, \tag{B.14}$$

we find the minimal solution for the model parameters by setting the gradient $\nabla_{\boldsymbol{w},b}\mathrm{MSE}$ to zero:

$$\left\langle (y - f) \frac{\partial}{\partial b} y \right\rangle = 0 \tag{B.15}$$

$$\left\langle (y - f) \frac{\partial}{\partial w_i} y \right\rangle = 0 \quad , \forall i \,. \tag{B.16}$$

This gives

$$\langle y \rangle = \langle f \rangle \tag{B.17}$$

$$\langle y \boldsymbol{x} \rangle = \langle f \boldsymbol{x} \rangle \,. \tag{B.18}$$

inserting the definition of y, one can solve for the parameters

$$b_0 = \langle f \rangle - \boldsymbol{c}_{fx} \widehat{C}_{xx}^{-1} \langle \boldsymbol{x} \rangle \tag{B.19}$$

$$\boldsymbol{w}_0 = \widehat{C}_{\mathrm{xx}}^{-1} \boldsymbol{c}_{\mathrm{fx}} \,, \tag{B.20}$$

where we defined

$$\widehat{C}_{\mathrm{xx}} \equiv \left\langle (\boldsymbol{x} - \langle \boldsymbol{x} \rangle) (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^T \right\rangle \tag{B.21}$$

$$\boldsymbol{c}_{\mathrm{fx}} \equiv \langle (f - \langle f \rangle) (\boldsymbol{x} - \langle \boldsymbol{x} \rangle) \rangle \,. \tag{B.22}$$

Using a similar approach, we now show that any vector that has the the same direction as $\boldsymbol{w}_0$ also maximizes the correlation.

From the definition of the Pearson correlation

$$\rho\left(y, f\right) \equiv \frac{\left\langle \left(y - \langle y \rangle\right)\left(f - \langle f \rangle\right)\right\rangle}{\sqrt{\left\langle \left(y - \langle y \rangle\right)^2 \right\rangle \left\langle \left(f - \langle f \rangle\right)^2 \right\rangle}} \tag{B.23}$$

we get for $\partial / \partial w_i \rho$

$$\frac{\partial \rho}{\partial w_i} = \frac{1}{\sigma_\mathrm{f} \sigma_\mathrm{y}^2} \left[\left\langle f \left(x_i - \langle x_i \rangle\right) \right\rangle \sigma_\mathrm{y}^2 - \left\langle y \left(f - \langle f \rangle\right) \right\rangle \left\langle y \left(x_i - \langle x_i \rangle\right) \right\rangle \right] = 0 \,. \tag{B.24}$$

From this equation, we get

$$\sigma_\mathrm{y}^2 \boldsymbol{c}_\mathrm{fx} = c_\mathrm{yf} \boldsymbol{c}_\mathrm{yx} \tag{B.25}$$

where, as for $\boldsymbol{c}_\mathrm{fx}$, c_yf and $\boldsymbol{c}_\mathrm{yx}$ denote the covariances between the corresponding variables. Expressing y via the linear model, this becomes

$$\left(\boldsymbol{w}^T \widehat{C}_\mathrm{xx} \boldsymbol{x} \right) \boldsymbol{c}_\mathrm{fx} = \left(\boldsymbol{w}^T \boldsymbol{c}_\mathrm{fx} \right) \widehat{C}_\mathrm{xx} \boldsymbol{w} \,. \tag{B.26}$$

It is straightforward to check that using $\boldsymbol{w}_0 = \lambda \widehat{C}_\mathrm{xx}^{-1} \boldsymbol{c}_\mathrm{fx}$ with an arbitrary scaling factor λ satisfies this equation, showing that maximal correlation is attained with linear coefficients that—up to an arbitrary scaling factor—are the same as for the minimal least squares solution.

Bibliography

[1] Kenneth S. Saladin, Stephen J. Sullivan, and Christina A. Gan. *Human Anatomy*. McGraw-Hill Education, New York, fifth edition, 2017. ISBN 978-0-07-340370-0.

[2] B Pakkenberg and H J Gundersen. Neocortical neuron number in humans: effect of sex and age. *The Journal of comparative neurology*, 384(2):312–320, jul 1997. ISSN 0021-9967 (Print).

[3] Mark F. Bear, Barry W. Connors, and Michael A. Paradiso. *Neuroscience: Exploring the Brain*. Lippincott Williams & Wilkins, Hagerstown, MD, third edition, 2007.

[4] Eric R. Kandel, James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and A. J. Hudspeth. *Priniples of Neuroscience*. McGraw-Hill, New, fifth edition, 2013. ISBN 978-0-07-139011-8.

[5] Misha V. Tsodyks and Henry Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences*, 94(2):719–723, jan 1997. ISSN 0027-8424. doi: 10.1073/PNAS.94.2.719.

[6] Branka Hrvoj-Mihic, Thibault Bienvenu, Lisa Stefanacci, Alysson R. Muotri, and Katerina Semendeferi. Evolution, development, and plasticity of the human brain: From molecules to bones. *Frontiers in Human Neuroscience*, 0(OCT):707, oct 2013. ISSN 16625161. doi: 10.3389/FNHUM.2013.00707/BIBTEX.

[7] J L Franklin, D J Fickbohm, and A L Willard. Long-term regulation of neuronal calcium currents by prolonged changes of membrane potential. *Journal of Neuroscience*, 12(5):1726–1735, 1992. ISSN 02706474. doi: 10.1523/jneurosci.12-05-01726.1992.

[8] Gina Turrigiano, L. F. Abbott, and Eve Marder. Activity-Dependent Changes in the Intrinsic Properties of Cultured Neurons. *Science*, 264(5161):974–977, 1994. doi: 10.1126/SCIENCE.8178157.

[9] Jochen Triesch. Synergies Between Intrinsic and Synaptic Plasticity Mechanisms. *Neural Computation*, 19(4):885–909, 2007. doi: 10.1162/neco.2007.19.4.885.

[10] D. Markovic and Claudius Gros. Self-Organized Chaos through Polyhomeostatic Optimization. *Physical Review Letters*, 105(6):068702, aug 2010. doi: 10.1103/PhysRevLett.105.068702.

[11] Jonathan Cannon and Paul Miller. Stable Control of Firing Rate Mean and Variance by Dual Homeostatic Mechanisms. *The Journal of Mathematical Neuroscience*, 7(1): 1, 2017.

[12] Paul Miller and Jonathan Cannon. Combined mechanisms of neural firing rate homeostasis. *Biological Cybernetics*, 113(1-2):47–59, apr 2019. ISSN 14320770. doi: 10.1007/S00422-018-0768-8/FIGURES/7.

[13] Francis Crick. The recent excitement about neural networks. *Nature 1989 337:6203*, 337(6203):129–132, 1989. ISSN 1476-4687. doi: 10.1038/337129a0.

[14] Eric I. Knudsen. Supervised learning in the brain. *Journal of Neuroscience*, 14(7): 3985–3997, jul 1994. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.14-07-03985.1994.

[15] Yoshua Bengio. How Auto-Encoders Could Provide Credit Assignment in Deep Networks via Target Propagation. jul 2014.

[16] Timothy P. Lillicrap, Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience 2020 21:6*, 21 (6):335–346, apr 2020. ISSN 1471-0048. doi: 10.1038/s41583-020-0277-3.

[17] F. G. Barker. Phineas among the phrenologists: the American crowbar case and nineteenth-century theories of cerebral localization. *Journal of Neurosurgery*, 82(4): 672–682, apr 1995. ISSN 00223085. doi: 10.3171/JNS.1995.82.4.0672.

[18] D. E. Haines. Central Nervous System, Overview. *Encyclopedia of the Neurological Sciences*, pages 637–640, jan 2014. doi: 10.1016/B978-0-12-385157-4.01130-1.

[19] L. F. Haas. Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(1):9–9, jan 2003. ISSN 0022-3050. doi: 10.1136/JNNP.74.1.9.

[20] S R Cajal. *Comparative Study of the Sensory Areas of the Human Cortex*. Clark University, 1899.

[21] H Barbas. General Cortical and Special Prefrontal Connections: Principles from Structure to Function. *Annual Review of Neuroscience*, 38(1):269–289, jul 2015. doi: 10.1146/annurev-neuro-071714-033936.

[22] A L Hodgkin and A F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117 (4):500–544, aug 1952. ISSN 1469-7793.

[23] David A. Drachman. Do we have brain to spare? *Neurology*, 64(12):2004–2005, jun 2005. ISSN 0028-3878. doi: 10.1212/01.WNL.0000166914.38327.BB.

[24] Bruce P. Bean. The action potential in mammalian central neurons. *Nature Reviews Neuroscience 2007 8:6*, 8(6):451–465, jun 2007. ISSN 1471-0048. doi: 10.1038/nrn2148.

[25] Wulfram Gerstner and Werner M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, Cambridge, UK, first edition, 2002. ISBN 978-0-511-07660-2.

[26] Wulfram Gerstner, Werner M. Kistler, Richard Naud, and Liam Paninski. *Neuronal Dynamics. From single neurons to networks and models of cognition (online book)*. Cambridge University Press, first edition, 2014.

[27] Romain Brette. Philosophy of the Spike: Rate-Based vs. Spike-Based Theories of the Brain. *Frontiers in Systems Neuroscience*, 9:151, 2015. ISSN 1662-5137. doi: 10.3389/fnsys.2015.00151.

[28] E. D. Adrian and Yngve Zotterman. The impulses produced by sensory nerve-endings. *The Journal of Physiology*, 61(2):151–171, apr 1926. ISSN 1469-7793. doi: 10.1113/JPHYSIOL.1926.SP002281.

[29] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106, jan 1962. ISSN 14697793. doi: 10.1113/JPHYSIOL.1962.SP006837.

[30] E. V. Evarts. Relation of pyramidal tract activity to force exerted during voluntary movement. *Journal of Neurophysiology*, 31(1):14–27, 1968. ISSN 00223077. doi: 10.1152/JN.1968.31.1.14.

[31] Donald J. Crammond and John F. Kalaska. Prior information in motor and premotor cortex: Activity during the delay period and effect on pre-movement activity. *Journal of Neurophysiology*, 84(2):986–1005, 2000. ISSN 00223077. doi: 10.1152/JN.2000.84.2.986/ASSET/IMAGES/LARGE/9K0801172012.JPEG.

[32] C. van Vreeswijk and H. Sompolinsky. Chaotic Balanced State in a Model of Cortical Circuits. *Neural Computation*, 10(6):1321–1371, aug 1998. ISSN 0899-7667. doi: 10.1162/089976698300017214.

[33] N Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J Comput Neurosci*, 8:183–208, 2000.

[34] P Dayan and L F Abbott. *Theoretical Neuroscience*. MIT Press, Cambridge, MA, USA, first edition, 2001.

[35] A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, and J. T. Massey. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience*, 2(11):1527–1537, nov 1982. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.02-11-01527.1982.

[36] Kanaka Rajan, Christopher D D. Harvey, and David W W. Tank. Recurrent Network Models of Sequence Generation and Memory. *Neuron*, 90(1):128–142, apr 2016. ISSN 1097-4199. doi: 10.1016/J.NEURON.2016.02.009.

[37] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, 46:1–6, oct 2017. ISSN 0959-4388. doi: 10.1016/J.CONB.2017.06.003.

[38] Rodney J. Douglas and Kevan A.C. Martin. Recurrent neuronal circuits in the neocortex. *Current Biology*, 17(13):R496–R500, jul 2007. ISSN 0960-9822. doi: 10.1016/J.CUB.2007.04.024.

[39] R J Douglas and K A Martin. A functional microcircuit for cat visual cortex. *The Journal of Physiology*, 440(1):735–769, aug 1991. ISSN 1469-7793. doi: 10.1113/JPHYSIOL.1991.SP018733.

[40] Steven Henry Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, first edition, 1994.

[41] Philip Hartman. A lemma in the theory of structural stability of differential equations. *Proceedings of the American Mathematical Society*, 11(4):610–620, apr 1960. ISSN 0002-9939. doi: 10.1090/S0002-9939-1960-0121542-7.

[42] Peter E. Kloeden and Martin Rasmussen. *Nonautonomous Dynamical Systems*. American Mathematical Society, Providence, Rhode Island, 2011. ISBN 978-1-4704-1403-0. doi: 10.1090/surv/176.

[43] Celso Grebogi, Edward Ott, Steven Pelikan, and James A. Yorke. Strange attractors that are not chaotic. *Physica D: Nonlinear Phenomena*, 13(1-2):261–268, aug 1984. ISSN 0167-2789. doi: 10.1016/0167-2789(84)90282-3.

[44] I. Gohberg, M. A. Kaashoek, and J. Kos. The asymptotic behavior of the singular values of matrix powers and applications. *Linear Algebra and its Applications*, 245: 55–76, sep 1996. ISSN 0024-3795. doi: 10.1016/0024-3795(96)82454-2.

[45] John Guckenheimer and Philip Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer, New York, first edition, 1983. ISBN 978-1-4612-1140-2.

[46] M. R.S. Kulenović and Orlando Merino. Global bifurcation for discrete competitive systems in the plane. *Discrete & Continuous Dynamical Systems - B. 2009, Volume 12, Pages 133-149*, 12(1):133, may 2009. ISSN 15313492. doi: 10.3934/DCDSB.2009.12.133.

[47] European Mathematical Society. Routes to chaos. *https://encyclopediaofmath.org/wiki/Routes_to_chaos*, 2020.

[48] Xin Wang. Period-Doublings to Chaos in A Simple Neural Network: An Analytical Proof. *Complex Systems*, 5, 1996.

[49] Yves Pomeau and Paul Manneville. Intermittent transition to turbulence in dissipative dynamical systems. *Communications in Mathematical Physics 1980 74:2*, 74(2):189–197, 1980. ISSN 1432-0916. doi: 10.1007/BF01197757.

[50] Ling-Wei Kong, Huawei Fan, Celso Grebogi, and Ying-Cheng Lai. Emergence of transient chaos and intermittency in machine learning. *J.Phys.Complex*, 2:35014–35030, 2021. doi: 10.1088/2632-072X/ac0b00.

[51] Bernard Doyon, Bruno Cessac, Mathias Quoy, and Manuel Samuelides. Destabilization and Route to Chaos in Neural Networks with Random Connectivity. In *NIPS*, pages 549–555, 1992.

[52] Marco Sandri. Numerical calculation of Lyapunov exponents. *Math. J.*, 6, 1996.

[53] Charalampos Skokos. The Lyapunov Characteristic Exponents and Their Computation. *Lecture Notes in Physics*, page 63, 2008. doi: 10.1007/978-3-642-04458-8_2.

[54] Valery Osedelet. A multiplicative ergodic theorem. Characteristic Ljapunov exponents of dynamical systems. *Transactions of the Moscow Mathematical Society*, 19:179—-210, 1968.

[55] Artur Dabrowski. Estimation of the largest Lyapunov exponent from the perturbation vector and its derivative dot product. *Nonlinear Dynamics 2011 67:1*, 67(1):283–291, mar 2011. ISSN 1573-269X. doi: 10.1007/S11071-011-9977-6.

[56] Giancarlo Benettin, Luigi Galgani, Antonio Giorgilli, and Jean Marie Strelcyn. Lyapunov Characteristic Exponents for smooth dynamical systems and for hamiltonian systems; a method for computing all of them. Part 1: Theory. *Meccanica*, 15(1):9–20, 1980. ISSN 00256455. doi: 10.1007/BF02128236.

[57] G. Leuba and L. J. Garey. Comparison of neuronal and glial numerical density in primary and secondary visual cortex of man. *Experimental Brain Research 1989 77:1*, 77(1):31–38, aug 1989. ISSN 1432-1106. doi: 10.1007/BF00250564.

[58] Mountcastle VB. The columnar organization of the neocortex. *Brain : a journal of neurology*, 120 (Pt 4(4):701–722, 1997. ISSN 0006-8950. doi: 10.1093/BRAIN/120.4.701.

[59] Daniel P. Buxhoeveden and Manuel F. Casanova. The minicolumn hypothesis in neuroscience. *Brain*, 125(5):935–951, may 2002. ISSN 0006-8950. doi: 10.1093/BRAIN/AWF110.

[60] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, sep 2011.

[61] Arthur W. Toga, Kristi A. Clark, Paul M. Thompson, David W. Shattuck, and John Darrell Van Horn. Mapping the Human Connectome. *Neurosurgery*, 71(1):1–5, jul 2012. ISSN 0148-396X. doi: 10.1227/NEU.0B013E318258E9FF.

[62] Seung Wook Oh, Julie A. Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M. Henry, Marty T. Mortrud, Benjamin Ouellette, Thuc Nghi Nguyen, Staci A. Sorensen, Clifford R. Slaughterbeck, Wayne Wakeman, Yang Li, David Feng, Anh Ho, Eric Nicholas, Karla E. Hirokawa, Phillip Bohn, Kevin M. Joines, Hanchuan Peng, Michael J. Hawrylycz, John W. Phillips, John G. Hohmann, Paul Wohnoutka, Charles R. Gerfen, Christof Koch, Amy Bernard, Chinh Dang, Allan R. Jones, and Hongkui Zeng. A mesoscale connectome of the mouse brain. *Nature 2014 508:7495*, 508(7495):207–214, apr 2014. ISSN 1476-4687. doi: 10.1038/nature13186.

[63] Paul C Bressloff. Spatiotemporal dynamics of continuum neural fields. *Journal of Physics A: Mathematical and Theoretical*, 45(3):033001, dec 2011. ISSN 1751-8121. doi: 10.1088/1751-8113/45/3/033001.

[64] H. Sompolinsky, A. Crisanti, and H. J. Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259–262, jul 1988. ISSN 00319007. doi: 10.1103/PhysRevLett.61.259.

[65] B. Cessac. Increase in Complexity in Random Neural Networks. *Journal de Physique I*, 5(3):409–432, mar 1995. ISSN 1155-4304. doi: 10.1051/JP1:1995135.

[66] Olivier Moynot and Manuel Samuelides. Large deviations and mean-field theory for asymmetric random recurrent neural networks. *Probability Theory and Related Fields 2002 123:1*, 123(1):41–75, may 2002. ISSN 1432-2064. doi: 10.1007/S004400100182.

[67] Olivier D Faugeras, Jonathan D Touboul, and Bruno Cessac. A constructive mean-field analysis of multi population neural networks with random synaptic weights and stochastic inputs. *Frontiers in Computational Neuroscience*, 0(FEB):1, feb 2009. ISSN 1662-5188. doi: 10.3389/NEURO.10.001.2009.

[68] Kanaka Rajan, L. F. Abbott, and Haim Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, 82(1):011903, jul 2010. doi: 10.1103/PhysRevE.82.011903.

[69] Marc Massar and Serge Massar. Mean-field theory of echo state networks. *Physical Review E*, 87:42809, 2013. doi: 10.1103/PhysRevE.87.042809.

[70] Jannis Schuecker, Sven Goedeke, and Moritz Helias. Optimal Sequence Memory in Driven Random Networks. *Physical Review X*, 8(4):041029, nov 2018. doi: 10.1103/PhysRevX.8.041029.

[71] L. D. Landau and E. M. Lifshitz. *Course of Theoretical Physics Volume 5: Statistical Physics*. Oxford, second edition, 1969.

[72] Mantas Lukoševičius. A Practical Guide to Applying Echo State Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7700 LECTU:659–686, 2012. doi: 10.1007/978-3-642-35289-8_36.

[73] G. L. Shaw. Donald Hebb: The Organization of Behavior. *Brain Theory*, pages 231–233, 1986. doi: 10.1007/978-3-642-70911-1_15.

[74] Karl Zilles. Neuronal plasticity as an adaptive property of the central nervous system. *Annals of Anatomy - Anatomischer Anzeiger*, 174(5):383–391, oct 1992. ISSN 0940-9602. doi: 10.1016/S0940-9602(11)80255-4.

[75] G G Turrigiano, K R Leslie, N S Desai, L C Rutherford, and S B Nelson. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391(6670):892–896, feb 1998. ISSN 00280836. doi: 10.1038/36103.

[76] Guo Qiang Bi and Mu Ming Poo. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464–10472, dec 1998. ISSN 02706474. doi: 10.1523/jneurosci.18-24-10464.1998.

[77] Ami Citri and Robert C Malenka. Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. *Neuropsychopharmacology 2008 33:1*, 33(1):18–41, aug 2007. ISSN 1740-634X. doi: 10.1038/sj.npp.1301559.

[78] Walter Bradford Cannon. Physiological regulation of normal states: some tentative postulates concerning biological homeostatics. In *A Charles Richet: ses Amis, ses Collegues, ses Eleves*. 1926.

[79] Walter Bradford Cannon. *The Wisdom of the Body*. Norton & Company, 1932.

[80] G G Turrigiano. Too Many Cooks? Intrinsic and Synaptic Homeostatic Mechanisms in Cortical Circuit Refinement. *Annual Review of Neuroscience*, 2011.

[81] Karl Johan Aström and Richard M. Murray. *Feedback Systems. An Introduction for Scientists and Engineers*. second edition, 2020.

[82] Attwell D and Laughlin SB. An energy budget for signaling in the grey matter of the brain. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, 21(10):1133–1145, 2001. ISSN 0271-678X. doi: 10.1097/00004647-200110000-00001.

[83] A J Bell and T J Sejnowski. An Information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[84] Fabian Schubert and Claudius Gros. Local Homeostatic Regulation of the Spectral Radius of Echo-State Networks. *Frontiers in Computational Neuroscience*, 0:12, feb 2021. ISSN 1662-5188. doi: 10.3389/FNCOM.2021.587721.

[85] A Lazar, G Pipa, and J Triesch. SORN: a self-organizing recurrent neural network. *Frontiers in Computational Neuroscience*, 3, 2009.

[86] Michiel W.H. Remme and Wytse J. Wadman. Homeostatic scaling of excitability in recurrent neural networks. *PLOS Computational Biology*, 8(5):1002494, may 2012. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002494.

[87] Friedemann Zenke, Guillaume Hennequin, and Wulfram Gerstner. Synaptic Plasticity in Neural Networks Needs Homeostasis with a Fast Rate Detector. *PLOS Computational Biology*, 9(11):1003330, nov 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003330.

[88] F Effenberger and J Jost. Self-Organization in Balanced State Networks by STDP and Homeostatic Plasticity. *PLOS Computational Biology*, 2015.

[89] Daniel Miner and Jochen Triesch. Plasticity-Driven Self-Organization under Topological Constraints Accounts for Non-random Features of Cortical Synaptic Wiring. *PLOS Computational Biology*, 12(2):e1004759, feb 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004759.

[90] Jochen J. Steil. Online reservoir adaptation by intrinsic plasticity for backpropagation-decorrelation and echo state learning. *Neural Networks*, 20(3):353–364, apr 2007. ISSN 08936080. doi: 10.1016/j.neunet.2007.04.011.

[91] Benjamin Schrauwen, Lars Buesing, and Robert Legenstein. On Computational Power and the Order-Chaos Phase Transition in Reservoir Computing. pages 1425–1432, oct 2008.

[92] J Boedecker, O Obst, N M Mayer, and M Asada. Initialization and self-organized optimization of recurrent neural network connectivity. *HFSP Journal*, 3(5):340–349, oct 2009. doi: 10.2976/1.3240502.

[93] Terence Tao and Van Vu. Random matrices: the circular law. *Communications in Contemporary Mathematics*, 10(02):261–307, 2008.

[94] Jérémie Barral and Alex D'Reyes. Synaptic scaling rule preserves excitatory-inhibitory balance and salient neuronal network dynamics. *Nature Neuroscience*, 19(12):1690–1696, dec 2016. ISSN 15461726. doi: 10.1038/nn.4415.

[95] H Jaeger. The "echo state" approach to analysing and training recurrent neural networks. GMD Report 148, GMD - German National Research Institute for Computer Science, 2001.

[96] Herbert Jaeger. Adaptive nonlinear system identification with echo state networks. *Advances in Neural Information Processing Systems*, 15:593—-600, 2002.

[97] Danko Nikolić, Stefan Häusler, Wolf Singer, and Wolfgang Maass. Distributed fading memory for stimulus properties in the primary visual cortex. *PLOS Biology*, 7(12): e1000260, 2009.

[98] Xavier Hinaut, Florian Lance, Colas Droin, Maxime Petit, Gregoire Pointeau, and Peter Ford Dominey. Corticostriatal response selection in sentence production: Insights from neural network simulation with reservoir computing. *Brain and Language*, 150: 54–68, nov 2015. ISSN 10902155. doi: 10.1016/j.bandl.2015.08.002.

[99] Pierre Enel, Emmanuel Procyk, René Quilodran, and Peter Ford Dominey. Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. *PLOS Computational Biology*, 12(6):e1004967, jun 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004967.

[100] Lorenzo Livi, Filippo Maria Bianchi, and Cesare Alippi. Determination of the edge of criticality in echo state networks through Fisher information maximization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(3):706–717, 2018.

[101] Yann Sweeney, Jeanette Hellgren Kotaleski, and Matthias H. Hennig. A Diffusive Homeostatic Signal Maintains Neural Heterogeneity and Responsiveness in Cortical Networks. *PLOS Computational Biology*, 11(7):e1004389, jul 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004389.

[102] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature 1986 323:6088*, 323(6088):533–536, 1986. ISSN 1476-4687. doi: 10.1038/323533a0.

[103] A. J. Robinson and Frank Fallside. The Utility Driven Dynamic Error Propagation Network. Technical report, Engineering Department, Cambridge University, Cambridge, UK, 1987.

[104] Paul J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4):339–356, jan 1988. ISSN 0893-6080. doi: 10.1016/0893-6080(88)90007-X.

[105] Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1):1–10, nov 2016. ISSN 20411723. doi: 10.1038/ncomms13276.

[106] Wolfgang Maass and Henry Markram. On the computational power of circuits of spiking neurons. *Journal of Computer and System Sciences*, 69(4):593–616, 2004. ISSN 00220000. doi: 10.1016/j.jcss.2004.04.001.

[107] Benjamin Schrauwen, David Verstraeten, and Jan Campenhout. An overview of reservoir computing: Theory, applications and implementations. In *Proceedings of the 15th European Sympsosium on Artificial Neural Networks*, pages 471–482, 2007.

[108] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2009.03.005.

[109] Arnaud Rachez and Masafumi Hagiwara. Augmented Echo State Networks with a feature layer and a nonlinear readout. *Proceedings of the International Joint Conference on Neural Networks*, 2012. doi: 10.1109/IJCNN.2012.6252505.

[110] Abdelkerim Souahlia, Ammar Belatreche, Abdelkader Benyettou, and Kevin Curran. An experimental evaluation of echo state network for colour image segmentation. *Proceedings of the International Joint Conference on Neural Networks*, 2016-Octob: 1143–1150, oct 2016. doi: 10.1109/IJCNN.2016.7727326.

[111] Michael Buehner and Peter Young. A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, 17(3):820–824, may 2006. doi: 10.1109/TNN.2006.872357.

[112] Gandhi Manjunath and Herbert Jaeger. Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. *Neural computation*, 25(3):671–696, 2013.

[113] Gilles Wainrib and Mathieu N. Galtier. A local Echo State Property through the largest Lyapunov exponent. *Neural Networks*, 76:39–45, apr 2016. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2015.12.013.

[114] V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, apr 1967. doi: 10.1070/SM1967V001N04ABEH001994.

[115] Marion Wardermann and Jochen Steil. Intrinsic plasticity for reservoir learning algorithms. In *ESANN 2007 Proceedings - 15th European Symposium on Artificial Neural Networks*, pages 513–518, 2007. ISBN 2930307099.

[116] Nils Bertschinger and Thomas Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, jul 2004. ISSN 08997667. doi: 10.1162/089976604323057443.

[117] Nicolas Schweighofer, Kenji Doya, Hidekazu Fukai, Jean Vianney Chiron, Tetsuya Furukawa, and Mitsuo Kawato. Chaos may enhance information transmission in the inferior olive. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4655, mar 2004. doi: 10.1073/PNAS.0305966101.

[118] Robert Legenstein and Wolfgang Maass. What makes a dynamical system computationally powerful? *New Directions in Statistical Signal Processing: From Systems to Brains*, 2005.

[119] Joschka Boedecker, Oliver Obst, Joseph T. Lizier, N. Michael Mayer, and Minoru Asada. Information processing in echo state networks at the edge of chaos. *Theory in Biosciences 2011 131:3*, 131(3):205–213, dec 2012. ISSN 1611-7530. doi: 10.1007/S12064-011-0146-8.

[120] L. Molgedey, J. Schuchhardt, and H. G. Schuster. Suppressing chaos in neural networks by noise. *Physical Review Letters*, 69(26):3717, dec 1992. doi: 10.1103/PhysRevLett.69.3717.

[121] Benjamin Schrauwen, Marion Wardermann, David Verstraeten, Jochen J Steil, and Dirk Stroobandt. Improving reservoirs using intrinsic plasticity. *Neurocomputing*, 71 (7-9):1159–1171, mar 2008. doi: 10.1016/j.neucom.2007.12.020.

[122] H Jaeger. Short Term Memory in Echo State Networks. GMD Report 152, Fraunhofer Institute for Autonomous Intelligent Systems, 2002.

[123] N Spruston. Pyramidal neurons: dendritic structure and synaptic integration. *Nature Reviews Neuroscience*, 9(3):206–221, mar 2008. doi: 10.1038/nrn2286.

[124] K Rajan and L F Abbott. Eigenvalue Spectra of Random Matrices for Neural Networks. *Physical Review Letters*, 97(18), nov 2006. doi: 10.1103/physrevlett.97.188104.

[125] Tom Binzegger, Rodney J. Douglas, and Kevan A.C. Martin. A quantitative map of the circuit of cat primary visual cortex. *Journal of Neuroscience*, 24(39):8441–8453, sep 2004. ISSN 02706474. doi: 10.1523/JNEUROSCI.1400-04.2004.

[126] W Martin Usrey and R Clay Reid. Synchronous Activity in the Visual System. *Annual Review of Physiology*, 61(1):435–456, 1999. ISSN 0066-4278. doi: 10.1146/annurev.physiol.61.1.435.

[127] Emilio Salinas and Terrence J Sejnowski. Correlated neuronal activity and the flow of neural information, aug 2001. ISSN 14710048.

[128] P Földiak. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 1990.

[129] Bryan Tripp and Chris Eliasmith. Neural Populations Can Induce Reliable Postsynaptic Currents without Observable Spike Rate Changes or Precise Spike Timing. *Cerebral Cortex*, 17(8):1830–1840, aug 2007. ISSN 1047-3211. doi: 10.1093/CERCOR/BHL092.

[130] Alexander S. Ecker, Philipp Berens, Georgios A. Keliris, Matthias Bethge, Nikos K. Logothetis, and Andreas S. Tolias. Decorrelated neuronal firing in cortical microcircuits. *Science*, 327(5965):584–587, jan 2010. ISSN 00368075. doi: 10.1126/SCIENCE.1179867/SUPPL_FILE/ECKER.SOM.PDF.

[131] Tom Tetzlaff, Moritz Helias, Gaute T. Einevoll, and Markus Diesmann. Decorrelation of Neural-Network Activity by Inhibitory Feedback. *PLOS Computational Biology*, 8 (8):e1002596, aug 2012. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1002596.

[132] Alberto Bernacchia and Xiao Jing Wang. Decorrelation by Recurrent Inhibition in Heterogeneous Neural Circuits. *Neural Computation*, 25(7):1732–1767, jul 2013. ISSN 0899-7667. doi: 10.1162/NECO_A_00451.

[133] Stuart D. Wick, Martin T. Wiechert, Rainer W. Friedrich, and Hermann Riecke. Pattern orthogonalization via channel decorrelation by adaptive networks. *Journal of Computational Neuroscience*, 28(1):29–45, feb 2010. ISSN 09295313. doi: 10.1007/S10827-009-0183-1/FIGURES/11.

[134] Keith B. Hengen, Alejandro Torrado Pacheco, James N. McGregor, Stephen D. Van Hooser, and Gina G. Turrigiano. Neuronal Firing Rate Homeostasis Is Inhibited by Sleep and Promoted by Wake. *Cell*, 165(1):180–191, mar 2016. ISSN 10974172. doi: 10.1016/J.CELL.2016.01.046/ATTACHMENT/9623821D-0B2E-4E0D-AA8E-F8337D03261C/MMC1.PDF.

[135] Andreas Frick and Daniel Johnston. Plasticity of dendritic excitability. *Journal of Neurobiology*, 64(1):100–115, jul 2005. ISSN 1097-4695. doi: 10.1002/NEU.20148.

[136] Jinhyun Kim, Sung Cherl Jung, Ann M. Clemens, Ronald S. Petralia, and Dax A. Hoffman. Regulation of Dendritic Excitability by Activity-Dependent Trafficking of the A-Type K+ Channel Subunit Kv4.2 in Hippocampal Neurons. *Neuron*, 54(6):933–947, jun 2007. ISSN 08966273. doi: 10.1016/J.NEURON.2007.05.026/ATTACHMENT/D07C0C03-5A53-467B-B7AB-165BB7A199C9/MMC3.MOV.

[137] Rebecca S. Hammond, Lin Lin, Michael S. Sidorov, Andrew M. Wikenheiser, and Dax A. Hoffman. Protein Kinase A Mediates Activity-Dependent Kv4.2 Channel Trafficking. *Journal of Neuroscience*, 28(30):7513–7519, jul 2008. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1951-08.2008.

[138] Na Chen, Xin Chen, and Jin Hui Wang. Homeostasis established by coordination of subcellular compartment plasticity improves spike encoding. *Journal of cell science*, 121(Pt 17):2961–2971, sep 2008. ISSN 0021-9533. doi: 10.1242/JCS.022368.

[139] Rishikesh Narayanan and Daniel Johnston. Functional maps within a single neuron. *Journal of neurophysiology*, 108(9):2343–2351, nov 2012. ISSN 1522-1598. doi: 10.1152/JN.00530.2012.

[140] Gina G. Turrigiano. The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, 135(3):422–435, oct 2008. ISSN 1097-4172. doi: 10.1016/J.CELL.2008.10.008.

[141] Frank Rosenblatt. The Perceptron–a perceiving and recognizing automaton. Technical report, Cornell Aeronautical Laboratory, 1957.

[142] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969.

[143] Stephen Grossberg. Competitive Learning: From Interactive Activation to Adaptive Resonance. *Cognitive Science*, 11(1):23–63, jan 1987. ISSN 1551-6709. doi: 10.1111/J.1551-6708.1987.TB00862.X.

[144] Y Bengio, Dong-Hyun Lee, Jorg Bornschein, and Zhouhan Lin. Towards Biologically Plausible Deep Learning. 2015.

[145] Timothy P. Lillicrap and Stephen H. Scott. Preference distributions of primary motor cortex neurons reflect control solutions optimized for limb biomechanics. *Neuron*, 77 (1):168–179, jan 2013. ISSN 1097-4199. doi: 10.1016/J.NEURON.2012.10.041.

[146] Seyed Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):e1003915, nov 2014. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1003915.

[147] Daniel L.K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, feb 2016. ISSN 1546-1726. doi: 10.1038/NN.4244.

[148] Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, may 2018. ISSN 1097-4199. doi: 10.1016/J.NEURON.2018.03.044.

[149] Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10(SEP): 94, sep 2016. ISSN 16625188. doi: 10.3389/FNCOM.2016.00094/BIBTEX.

[150] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11:24, may 2017. ISSN 16625188. doi: 10.3389/FNCOM.2017.00024/BIBTEX.

[151] J Guerguiev, T P Lillicrap, and B A Richards. Towards deep learning with segregated dendrites. *eLife*, 6, dec 2017. doi: 10.7554/elife.22901.

[152] Leopoldo Petreanu, Tianyi Mao, Scott M. Sternson, and Karel Svoboda. The subcellular organization of neocortical excitatory connections. *Nature 2009 457:7233*, 457 (7233):1142–1145, feb 2009. ISSN 1476-4687. doi: 10.1038/nature07709.

[153] Matthew Larkum. A cellular mechanism for cortical associations: An organizing principle for the cerebral cortex, mar 2013. ISSN 01662236.

[154] R Urbanczik and W Senn. Learning by the Dendritic Prediction of Somatic Spiking. *Neuron*, 81(3):521–528, feb 2014. doi: 10.1016/j.neuron.2013.11.030.

[155] Arild Nøkland. Direct Feedback Alignment Provides Learning in Deep Neural Networks. *Advances in Neural Information Processing Systems*, pages 1045–1053, sep 2016. ISSN 10495258.

[156] João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[157] Dong Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference Target Propagation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9284:498–515, 2015. ISSN 16113349. doi: 10.1007/978-3-319-23528-8_31.

[158] Alexander G. Ororbia and Ankur Mali. Biologically Motivated Algorithms for Propagating Local Target Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4651–4658, jul 2019. ISSN 2374-3468. doi: 10.1609/AAAI.V33I01.33014651.

[159] Fabian Schubert and Claudius Gros. Nonlinear Dendritic Coincidence Detection for Supervised Learning. *Frontiers in Computational Neuroscience*, 15:72, aug 2021. ISSN 16625188. doi: 10.3389/FNCOM.2021.718020/BIBTEX.

[160] Josef P. Rauschecker. Auditory and visual cortex of primates: a comparison of two sensory systems. *The European journal of neuroscience*, 41(5):579, mar 2015. ISSN 14609568. doi: 10.1111/EJN.12844.

[161] Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 1(1):1, 1991. ISSN 1047-3211. doi: 10.1093/CERCOR/1.1.1-A.

[162] Dwight J. Kravitz, Kadharbatcha S. Saleem, Chris I. Baker, Leslie G. Ungerleider, and Mortimer Mishkin. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17(1):26–49, jan 2013. ISSN 1364-6613. doi: 10.1016/J.TICS.2012.10.011.

[163] Charles G. Gross. Genealogy of the "grandmother cell". *Neuroscientist*, 8(5):512–518, jun 2002. ISSN 10738584. doi: 10.1177/107385802237175.

[164] Keiji Tanaka. Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience*, 19:109–139, nov 1996. ISSN 0147006X. doi: 10.1146/ANNUREV.NE.19.030196.000545.

[165] Daniel J. Felleman, A. Burkhalter, and David C. Van Essen. Cortical connections of areas V3 and VP of macaque monkey extrastriate visual cortex. *Journal of Comparative Neurology*, 379:21–47, 1997.

[166] Claus C. Hilgetag and Alexandros Goulas. 'Hierarchy' in the organization of brain networks. *Philosophical Transactions of the Royal Society B*, 375(1796), apr 2020. ISSN 14712970. doi: 10.1098/RSTB.2019.0319.

[167] Robert D. Brandt and Feng Lin. Supervised learning in neural networks without feedback network. *IEEE International Symposium on Intelligent Control - Proceedings*, pages 86–90, 1996. doi: 10.1109/ISIC.1996.556182.

[168] Kenneth D. Harris. Stability of the fittest: organizing learning through retroaxonal signals. *Trends in Neurosciences*, 31(3):130–136, mar 2008. ISSN 0166-2236. doi: 10.1016/J.TINS.2007.12.002.

[169] Randall C. O'Reilly. Biologically Plausible Error-Driven Learning Using Local Activation Differences: The Generalized Recirculation Algorithm. *Neural Computation*, 8 (5):895–938, jul 1996. ISSN 0899-7667. doi: 10.1162/NECO.1996.8.5.895.

[170] Xiaohui Xie and H. Sebastian Seung. Equivalence of Backpropagation and Contrastive Hebbian Learning in a Layered Network. *Neural Computation*, 15(2):441–454, feb 2003. ISSN 0899-7667. doi: 10.1162/089976603762552988.

[171] Charlotte Frenkel, Martin Lefebvre, and David Bol. Learning Without Feedback: Fixed Random Learning Signals Allow for Feedforward Training of Deep Neural Networks. *Frontiers in Neuroscience*, 15:20, feb 2021. ISSN 1662453X. doi: 10.3389/FNINS.2021.629892/BIBTEX.

[172] Arild Nøkland and Lars Hiller Eidnes. Training Neural Networks with Local Error Signals. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4839–4850. PMLR, 2019.

[173] Theodore H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep convolutional networks. dec 2018. ISSN 2331-8422.

[174] Brian Crafton, Abhinav Parihar, Evan Gebhardt, and Arijit Raychowdhury. Direct feedback alignment with sparse connections for local learning. *Frontiers in Neuroscience*, 13(MAY):525, 2019. ISSN 1662453X. doi: 10.3389/FNINS.2019.00525/BIBTEX.

[175] Yann Le Cun. Learning Process in an Asymmetric Threshold Network. *Disordered Systems and Biological Organization*, pages 233–240, 1986. doi: 10.1007/978-3-642-82657-3_24.

[176] Charles D. Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience 2013 14:5*, 14(5):350–363, apr 2013. ISSN 1471-0048. doi: 10.1038/nrn3476.

[177] Aris Fiser, David Mahringer, Hassana K. Oyibo, Anders V. Petersen, Marcus Leinweber, and Georg B. Keller. Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience 2016 19:12*, 19(12):1658–1664, sep 2016. ISSN 1546-1726. doi: 10.1038/nn.4385.

[178] Marcus Leinweber, Daniel R. Ward, Jan M. Sobczak, Alexander Attinger, and Georg B. Keller. A Sensorimotor Circuit in Mouse Cortex for Visual Flow Predictions. *Neuron*, 95(6):1420–1432.e5, sep 2017. ISSN 10974199. doi: 10.1016/J.NEURON.2017.08.036/ATTACHMENT/E29E8E8B-0B74-4246-8522-446C236A909C/MMC3.MP4.

[179] Etay Hay, Sean Hill, Felix Schürmann, Henry Markram, and Idan Segev. Models of Neocortical Layer 5b Pyramidal Cells Capturing a Wide Range of Dendritic and Perisomatic Active Properties. *PLOS Computational Biology*, 7(7):e1002107, jul 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002107.

[180] Srikanth Ramaswamy and Henry Markram. Anatomy and physiology of the thick-tufted layer 5 pyramidal neuron, jun 2015. ISSN 16625102.

[181] A S Shai, C A Anastassiou, M E Larkum, and C Koch. Physiology of Layer 5 Pyramidal Neurons in Mouse Primary Visual Cortex: Coincidence Detection through Bursting. *PLOS Computational Biology*, 11(3), 2015.

[182] Nelson Spruston, Yitzhak Schiller, Greg Stuart, and Bert Sakmann. Activity-dependent action potential invasion and calcium influx into hippocampal CA1 dendrites. *Science*, 268(5208):297–300, apr 1995. ISSN 00368075. doi: 10.1126/science.7716524.

[183] M. Häusser, N. Spruston, and G. J. Stuart. Diversity and dynamics of dendritic signaling, oct 2000. ISSN 00368075.

[184] Tiago Branco and Michael Häusser. Synaptic Integration Gradients in Single Cortical Pyramidal Cell Dendrites. *Neuron*, 69(5):885–892, mar 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.02.006.

[185] Nelson Spruston, David B. Jaffe, and Daniel Johnston. Dendritic attenuation of synaptic potentials and currents: the role of passive membrane properties. *Trends in Neurosciences*, 17(4):161–166, jan 1994. ISSN 0166-2236. doi: 10.1016/0166-2236(94)90094-9.

[186] Stephen R. Williams and Greg J. Stuart. Dependence of EPSP efficacy on synapse location in neocortical pyramidal neurons. *Science*, 295(5561):1907–1910, mar 2002. ISSN 00368075. doi: 10.1126/SCIENCE.1067903/SUPPL_FILE/1067903S2_THUMB.GIF.

[187] H. T. Chang. Dendritic Potential of Cortical Neurons produced by Direct Electrical Stimulation of the Cerebral Cortex. *Journal of Neurophysiology*, 14(1):1–21, jan 1951. ISSN 00223077. doi: 10.1152/JN.1951.14.1.1.

[188] M. W. Spratling. Cortical region interactions and the functional role of apical dendrites. *Behavioral and cognitive neuroscience reviews*, 1(3):219–228, may 2002. ISSN 15345823. doi: 10.1177/1534582302001003003.

[189] David LaBerge and Ray Kasevich. The apical dendrite theory of consciousness. *Neural Networks*, 20(9):1004–1020, nov 2007. ISSN 0893-6080. doi: 10.1016/J.NEUNET.2007.09.006.

[190] Jackie Schiller, Yitzhak Schiller, Greg Stuart, and Bert Sakmann. Calcium action potentials restricted to distal apical dendrites of rat neocortical pyramidal neurons. *The Journal of Physiology*, 505(Pt 3):605, dec 1997. ISSN 00223751. doi: 10.1111/J.1469-7793.1997.605BA.X.

[191] Stephen R. Williams and Greg J. Stuart. Mechanisms and consequences of action potential burst firing in rat neocortical pyramidal neurons. *The Journal of physiology*, 521 Pt 2(Pt 2):467–482, dec 1999. ISSN 0022-3751. doi: 10.1111/J.1469-7793.1999.00467.X.

[192] Matthew E. Larkum and J. Julius Zhu. Signaling of Layer 1 and Whisker-Evoked Ca2+ and Na+ Action Potentials in Distal and Terminal Dendrites of Rat Neocortical Pyramidal Neurons In Vitro and In Vivo. *Journal of Neuroscience*, 22(16):6991–7005, aug 2002. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.22-16-06991.2002.

[193] M. E. Larkum, J. J. Zhu, and B. Sakmann. Dendritic mechanisms underlying the coupling of the dendritic with the axonal action potential initiation zone of adult rat layer 5 pyramidal neurons. *The Journal of Physiology*, 533(Pt 2):447, jun 2001. ISSN 00223751. doi: 10.1111/J.1469-7793.2001.0447A.X.

[194] B. Gustafsson, H. Wigstrom, W. C. Abraham, and Y. Y. Huang. Long-term potentiation in the hippocampus using depolarizing current pulses as the conditioning stimulus to single volley synaptic potentials. *Journal of Neuroscience*, 7(3):774–780, 1987. ISSN 02706474. doi: 10.1523/jneurosci.07-03-00774.1987.

[195] Dominique Debanne, Beat H. Gähwiler, and Scott M. Thompson. Asynchronous pre- and postsynaptic activity induces associative long-term depression in area CA1 of the rat hippocampus in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 91(3):1148–1152, feb 1994. ISSN 00278424. doi: 10.1073/pnas.91.3.1148.

[196] Henry Markram, Joachim Lübke, Michael Frotscher, and Bert Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297): 213–215, jan 1997. ISSN 00368075. doi: 10.1126/science.275.5297.213.

[197] Xiaohui Xie and H. Sebastian Seung. Learning in neural networks by reinforcement of irregular spiking. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 69(4):10, apr 2004. ISSN 1063651X. doi: 10.1103/PHYSREVE.69.041909/FIGURES/7/MEDIUM.

[198] Robert Legenstein, Dejan Pecevski, and Wolfgang Maass. A Learning Theory for Reward-Modulated Spike-Timing-Dependent Plasticity with Application to Biofeedback. *PLOS Computational Biology*, 4(10):e1000180, oct 2008. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1000180.

[199] Nicolas Frémaux, Henning Sprekeler, and Wulfram Gerstner. Reinforcement Learning Using a Continuous Time Actor-Critic Framework with Spiking Neurons. *PLOS Computational Biology*, 9(4):e1003024, apr 2013. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1003024.

[200] Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, 12:53, jul 2018. ISSN 16625110. doi: 10.3389/FNCIR.2018.00053/BIBTEX.

[201] E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48, jan 1982. ISSN 02706474. doi: 10.1523/jneurosci.02-01-00032.1982.

[202] Nathan Intrator and Leon N. Cooper. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5(1):3–17, jan 1992. ISSN 08936080. doi: 10.1016/S0893-6080(05)80003-6.

[203] R Linsker. From basic network principles to neural architecture: emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences of the United States of America*, 83(21):8390, 1986. doi: 10.1073/PNAS.83.21.8390.

[204] Pierre Yger and Matthieu Gilson. Models of Metaplasticity: A Review of Concepts. *Frontiers in Computational Neuroscience*, 9(November):138, nov 2015. ISSN 1662-5188. doi: 10.3389/FNCOM.2015.00138.

[205] C. Charles Law and Leon N. Cooper. Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory. *Proceedings of the National Academy of Sciences of the United States of America*, 91(16): 7797–7801, aug 1994. ISSN 00278424. doi: 10.1073/pnas.91.16.7797.

[206] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, New York, 2013. doi: 10.1007/978-1-4614-7138-7.

[207] Wolfgang Härdle and Leopold Simar. Canonical Correlation Analysis. In *Applied Multivariate Statistical Analysis*, pages 321–330. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. doi: 10.1007/978-3-540-72244-1_14.

[208] Tatsuya Haga and Tomoki Fukai. Dendritic processing of spontaneous neuronal sequences for single-trial learning. *Scientific Reports*, 8(1):15166, dec 2018. ISSN 20452322. doi: 10.1038/s41598-018-33513-9.

[209] Claudia Clopath, Lars Büsing, Eleni Vasilaki, and Wulfram Gerstner. Connectivity reflects coding: A model of voltage-based STDP with homeostasis. *Nature Neuroscience*, 13(3):344–352, mar 2010. ISSN 10976256. doi: 10.1038/nn.2479.

[210] Felix Weissenberger, Marcelo Matheus Gauy, Johannes Lengler, Florian Meier, and Angelika Steger. Voltage dependence of synaptic plasticity is essential for rate based learning with short stimuli. *Scientific Reports*, 8(1):4609, dec 2018. ISSN 20452322. doi: 10.1038/s41598-018-22781-0.

[211] Timothy O'Leary and David J.A. Wyllie. Neuronal homeostasis: time for a change? *The Journal of Physiology*, 589(20):4811–4826, oct 2011. ISSN 1469-7793. doi: 10.1113/JPHYSIOL.2011.210179.

[212] Steven L Brunton and J Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019. doi: 10.1017/9781108380690.

[213] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *https://doi.org/10.1007/BF02418571*, 30(none):175–193, jan 1906. ISSN 0001-5962. doi: 10.1007/BF02418571.

Acknowledgments

My thanks go to the many people that supported me over the course of my PhD. This work would not have been possible without them.

First of all, I would like to thank Prof. Claudius Gros for his scientific support and guidance. I am very grateful for the freedom that was given to me in following my research interests and for all the fruitful conversations we had over the years, often exceeding the scope of our academic work.

I would also like to thank Prof. Jochen Triesch as the second referee for this thesis.

I also want to thank all my friends and colleagues in Prof. Gros' group for the scientific and personal support that I received over the years: Hendrik Wernecke, Emanuele Varriale, Oren Neumann, Lukas Schneider, Tim Koglin, Carolin Roskothen, Frederike Kubandt and Michael Nowak.

My special thanks go to Anna Oertl, Karim Zantout, Arthur Scammell and Lukas Koehs for their scientific and emotional help and support—both within and outside my work.

Curriculum Vitae

Fabian Schubert

address	Rohrbachstr. 60 60389 Frankfurt am Main Germany
phone	+15781904107
e-mail	fschubert@itp.uni-frankfurt.de
birthday	06.11.1990
place of birth	Frankfurt Höchst
nationality	German

Education

2017 – 2022	PhD student in the group of Prof. Dr. Claudius Gros Institute for Theoretical Physics, Goethe University Frankfurt
2017	Master of Science in Physics, Goethe University Frankfurt Master's Thesis supervised by Prof. Dr. Jochen Triesch, Frankfurt Institute for Advanced Studies. Thesis title: "Diffusive Homeostasis in a Self-Organizing Recurrent Neural Network: Spatially dependent Interaction as a Determinant of Neural Activity and Plasticity"
2014	Bachelor of Science in Physics, Goethe University Frankfurt Bachelor's Thesis supervised by Prof. Dr. Claudius Gros, Institute for Theoretical Physics, Goethe University Frankfurt. Thesis title: "Studies on a Two-dimensional Spiking Neuron Model: Bifurcation and Excitability"
2010	university-entrance diploma, Augustinerschule Friedberg

Teaching

summer term 2021, winter term 2018	tutor "Complex Adaptive Dynamical Systems"
winter term 2020	tutor "Programming for Physicists"
summer term 2020, summer term 2018	tutor "Self-Organization: Theory and Simulations"
winter term 2019, winter term 2018	tutor "Advanced Introduction to C++, Scientific Computing and Machine Learning"
winter term 2017	tutor "Theoretical Physics 5: Statistical Mechanics and Thermodynamics"

Publications

F. Schubert, C. Gros.
Nonlinear Dendritic Coincidence Detection for Supervised Learning.
Frontiers in Computational Neuroscience, 2021.

F. Schubert, C. Gros.
Local Homeostatic Regulation of the Spectral Radius of Echo-State Networks.
Frontiers in Computational Neuroscience, 2021.

Other Scientific Work

F. Schubert, C. Gros. *Goethe Interactive COVID-19 Analyzer.*
`https://itp.uni-frankfurt.de/covid-19/`
Interface design and technical implementation of front-end and back-end.

Awards

2020	Best presentation award in computational neuroscience NEURONUS 2020 IBRO Neuroscience Forum

Talks

	Local Autonomous Online Regulation of Echo-State Networks:
2021	ICNAAM 2021 - Symposium on Mathematics of Neuro-Science, Technology and Engineering
2021	APS March Meeting 2021
2020	NEURONUS 2020 IBRO Neuroscience Forum
2020	Brain Criticality Meeting

Posters

	Nonlinear Dendritic Integration Increases Alignment of Basal and Apical Input under Hebbian Plasticity.
2021	Bernstein Conference 2021
	Local variance optimization for the autonomous regulation of echo state networks.
2020	Bernstein Conference 2020
2019	Interdisciplinary College, Günne, Germany
	A Continuous-Time Dynamical System Describing both Rate Encoding and Spiking Neurons.
2018	Analysis and Modeling of Complex Oscillatory Systems, Barcelona, Spain

Schools and Workshops

2019	Interdisciplinary College Günne, Germany
2017	Summerschool on Advanced Computational Neurosciences Göttingen, Germany