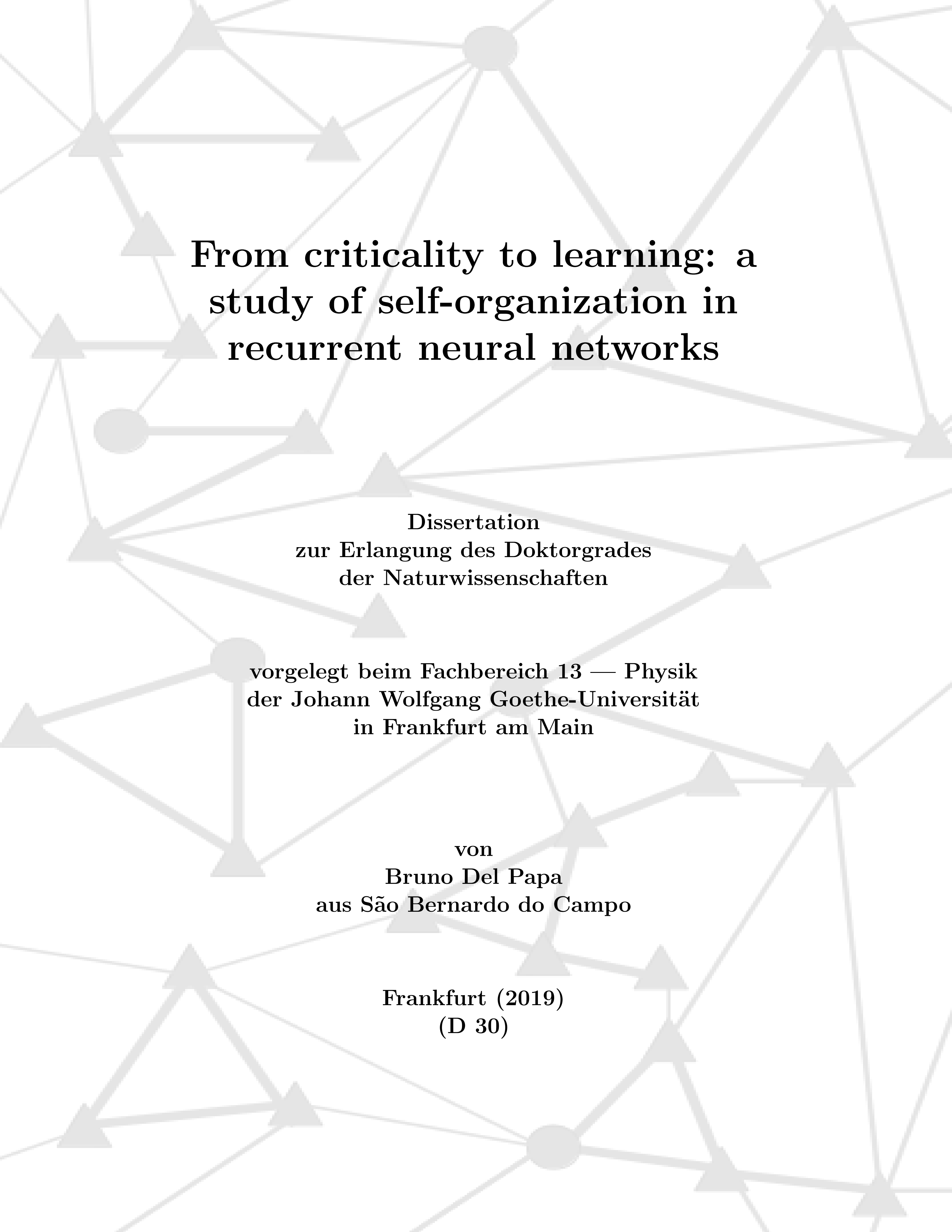


From criticality to learning: a study of self-organization in recurrent neural networks

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 13 — Physik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Bruno Del Papa
aus São Bernardo do Campo

Frankfurt (2019)
(D 30)

Cover: Schematic representation of a sparsely connected recurrent neural network with excitatory (triangles) and inhibitory (circles) neurons.

vom Fachbereich 13 der
Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:
Prof. Dr. Michael Lang

Gutachter:
Prof. Dr. Jochen Triesch
Dr. Viola Priesemann

Datum der Disputation: 28.10.2019

Contributions

This thesis is based on published and currently submitted manuscripts. The manuscripts have been accepted by or will be soon submitted to peer-reviewed journals, and were co-authored by Bruno Del Papa (BDP), Viola Priesemann (VP), Jochen Triesch (JT), Antonia Hufnagl (AN) and Florence Kleberg (FK). The following is a chronological overview of the individual contributions and the current state of each of those works, as of October 29, 2019.

Del Papa et al. (2017): Criticality meets learning: criticality signatures in a self-organizing recurrent neural network. The model presented in Chapter 3 and a substantial part of the results and conclusions appearing in Chapter 4 have been published in the journal PLOS ONE as a research article by BDP, VP, and JT.

- BDP, VP, and JT conceived and designed the work
- BDP wrote and did all the simulations, analyses, and visualizations
- BDP and VP designed the experiments on unstructured external input
- BDP and JT designed the experiments on structured external input and learning
- VP and JT supervised the work
- BDP wrote the first draft of the manuscript
- BDP, VP, and JT reviewed and edited the manuscript

Del Papa et al. (2019): Fading memory, plasticity, and criticality in recurrent networks. The model presented in Chapter 3, a small part of the results and discussions appearing in Chapter 4, and the fading memory analysis in Chapter 5 have been accepted as a book chapter for the book *The* *functional role of critical dynamics*, by Springer, and will soon be published. This book chapter was authored by BDP, VP, and JT.

- BDP and JT conceived and designed the manuscript
- BDP did all the simulations (adapted from Del Papa et al. (2017)), analyses, and visualizations
- JT proposed the study of the fading memory capacity and its relations to critical dynamics
- VP and JT supervised the work
- BDP wrote the first draft of the manuscript
- BDP, VP, and JT reviewed and edited the manuscript

Creative sentence generation in a deterministic Self-Organizing Recurrent Neural network. Part of the results from Chapter 5 regarding grammar and sentence learning and generation will be submitted as a research article by AH, BDP, FK, and JT. The manuscript is currently in preparation.

- AH, BDP, and JT conceived and designed the study
- AH and BDP wrote and did all the simulations, analyses, and visualizations
- AH designed the artificial dictionaries and performed the experiments on character-level grammar learning
- BDP and JT designed the experiments on the CHILDES dataset
- BDP wrote the simulations comparing SORNs and other recurrent neural network models
- JT supervised the work
- FK reviewed the text and edited the introduction and methods
- AH, BDP, FK, and JT are writing the first draft of the manuscript

Acknowledgements

It has been a critical adventure! Roughly four years ago, I moved to Frankfurt and began working on my PhD, from which this thesis is the most complete research summary. The thesis fails, however, to capture all the life-changing events that resulted directly or indirectly from this work, all the personal growth that followed, and all the indispensable assistance I have had from many people since it all started. Of course, only a brief mention here does not do everybody justice, but I hope it serves as a simple, but long lasting, thank you note.

First, the whole idea of doing a PhD would have been inconceivable if not for my parents, Hélia and Marcos. They have done whatever they could to support me in all endeavors I have chosen to pursue, including flying across the ocean to work on a relatively obscure topic in computational neuroscience. For as long as I can remember, they have made sure I could keep studying what I wanted, and most importantly, have always encouraged me to never stop learning and always stay curious. I have tried to follow their advice, and not only this thesis but all work I have produced so far is the result of their encouragement. Thank you, I am more grateful to you two than I may have ever explicitly said.

Second, I am very grateful to all the friends and colleagues I have met during these years, especially the ones who have truly believed the countless times I have claimed to be *almost* done with writing. I might one day forget the details of this piece of research, but I will not forget the people with whom I shared some of the best years of my life. Namely, I would like to thank Jagoda, Rosana, Alexandra, Kauê, the current and former trieschies (Lukas, Natalie, Florence, Alex, Diyuan, Charles, Max, and Christoph — who got me into SORNs in the first place), Bettina, Basti, Enrique, among many other wonderful persons I have met at FIAS, the MPI for Brain Research, and other not so random corners of the world. I probably would have to write a new book to describe everything I learned from you all and how it changed my life. Thank you for the company, the stories, the board games, the parties, Stammtisch and karaoke evenings, late night beers and, not least, for the occasional free food. You guys kept me sane during grad school!

Special thanks and double mentions go to my friends who have kindly translated to German the first sections of this thesis: Bettina (twice), Lukas, Bastian (twice), Natalie, Alex L, Sigrid, and Nicklas. You and a few others have made my (hopefully temporary) lack of German language skills less of a burden over the last years. Similarly, my stay in Frankfurt would have not been so successful if not for the countless times I received assistance from the international office at the MPI for Brain Research, particularly from Maren, who very patiently walked me through some obscure corners of German bureaucracy. Also in the MPI, I am thankful to Arjan and the other IMPRS people for accepting me into the graduate program, for the support during my first year, and for organizing the most enjoyable scientific retreats, exactly poised between ordered and chaotic dynamics.

Finally, this work would not have been possible without the support of my enthusiastic supervisor, Jochen Triesch, whose various scientific interests have given me the motivation to self-organize and learn new concepts even in relatively unfamiliar topics. I am also thankful to Viola Priesemann for critical discussions about criticality in neural circuits, which ended up shaping most results and conclusions of this thesis. Moreover, I am thankful to FIAS, MPI, and university staff who daily contribute to the maintenance of a great research atmosphere (in particular to the staff at Cafeteria Darwins in the Riedberg Campus, for what has become my research motto at lunchtime — “Eine gute pasta ist das Fundament allen Glücks”).

Bruno

Abstract

The brain is a large complex system which is remarkably good at maintaining stability under a wide range of input patterns and intensities. In addition, such a stable dynamical state is able to sustain essential functions, including the encoding of information about the external environment and storing memories. In order to succeed in these challenging tasks, neural circuits rely on a variety of plasticity mechanisms that act as self-organizational rules and regulate their dynamics. Based on toy models of self-organized criticality, this stable state has been proposed to be a phase transition point, poised between distinct types of unhealthy dynamics, in what has become known as the critical brain hypothesis. It is not yet known, however, if and how self-organization could drive biological neural networks towards a critical state while maintaining or improving their learning and memory functions.

Here, we investigate the emergence of criticality signatures in the form of neuronal avalanches due to self-organizational plasticity rules in a recurrent neural network. We show that power-law distributions of events, widely observed in experiments, arise from a combination of biologically inspired synaptic and homeostatic plasticity but are highly dependent on the external drive. Additionally, we describe how learning abilities and fading memory emerge and are improved by the same self-organizational processes. We finally propose an application of these enhanced functions, focusing on sequence and simple language learning tasks.

Taken together, our results suggest that the same self-organizational processes can be responsible for improving the brain's spatio-temporal learning abilities and memory capacity while also giving rise to criticality signatures under particular input conditions, thus proposing a novel link between such abilities and neuronal avalanches. Although criticality was not verified, the detailed study of self-organization towards critical dynamics further elucidates its potential emergence and functions in the brain.

Kurze Zusammenfassung

Obwohl das Gehirn diversen äußeren Reizen in verschiedensten Intensitäten ausgesetzt ist, zeigt es als komplexes dynamisches System eine bemerkenswerte Stabilität. Solche stabilen dynamischen Zustände erlauben die Aufrechterhaltung essenzieller Funktionen, insbesondere das Speichern von Erinnerungen oder die Kodierung der externen Welt in neuronalen Aktivitätsmustern. Um diese anspruchsvollen Aufgaben zu bewältigen, sind neuronale Schaltkreise auf verschiedene Plastizitätsmechanismen angewiesen, welche als selbst-organisierende Regeln die Netzwerkdynamik regulieren. Die Hypothese vom Gehirn als System am kritischen Punkt schlägt ausgehend von vereinfachten Modellen der selbst-organisierten Kritikalität vor, dass dieser stabile Zustand ein Phasenübergang zwischen verschiedenen Arten anormaler beziehungsweise ungesunder Dynamik ist. Jedoch ist noch unklar, ob und wie biologische neuronale Netzwerke durch Selbst-Organisation zu einem kritischen Punkt gelangen, während ihre Lern- und Gedächtnisfunktionen aufrechterhalten oder verbessert werden.

In dieser Arbeit untersuchen wir die Entstehung von Anzeichen für Kritikalität in Form von neuronalen Aktivitätslawinen, die durch selbst-organisierende Plastizitätsregeln in rekurrent verbundenen neuronalen Netzwerken auftreten. Wir zeigen, dass die experimentell beobachtbare Potenzgesetz-Verteilung der Aktivitätslawinen aus einer Kombination von biologisch inspirierten synaptischen und homöostatischen Plastizitätsregeln entstehen, aber stark von externem Input abhängig sind. Zusätzlich beschreiben wir, wie die Fähigkeit zu Lernen und zu Vergessen durch die gleichen selbst-organisierenden Prozesse sowohl entstehen als auch sich verbessern können. Abschließend zeigen wir mögliche Anwendungen dieser verbesserten Funktionen auf, mit einem Fokus auf dem Lernen von Sequenzen sowie einfachen Sprach-Lernaufgauben.

Alles in allem legen unsere Ergebnisse nahe, dass die gleichen selbst-organisierenden Prozesse sowohl für die Verbesserung von räumlich-zeitlich abhängigem Lernen als auch für die Gedächtniskapazität verantwortlich sind. Gleichzeitig rufen diese Prozesse Anzeichen für Kritikalität unter bestimmten Bedingungen für den Input hervor und regen eine neue Verknüpfung dieser Fähigkeiten des Gehirns mit neuronalen Aktivitätslawinen an. Obwohl diese Arbeit keine Kritikalität im Gehirn nachweisen kann, verdeutlicht die detaillierte Untersuchung der Selbstorganisation zur Kritikalität ihr potenzielles Auftreten und ihre Funktionen im Gehirn.

Zusammenfassung[1]

Selbstorganisation, Kritikalität und das Gehirn

Eines der anspruchvollsten Probleme der modernen Wissenschaft ist, die Funktionsweise des Gehirns zu verstehen. Als komplexes System aus Hunderten von Milliarden nichtlinearer Informationsverarbeitungseinheiten, die durch Hunderte von Billionen Synapsen miteinander verbunden sind, ähnelt die kollektive Dynamik des Gehirns vielen klassischen Systemen der statistischen Physik. Darüber hinaus passt sich dieses enorme System ständig dynamisch an, wodurch ein großes Repertoire an räumlich-zeitlichen Aktivitätsmustern offenbar wird. Man kennt mehrere Anpassungsmechanismen, die simultan wirken und die Gehirnaktivität regulieren, welche praktisch allen lebenswichtigen Funktionen zugrunde liegt, von grundlegenden Muskelbewegungen bis hin zu anspruchsvollen kognitiven Prozessen. Diese auftretenden neuronalen Phänomene finden unabhängig von äußerer Kontrolle statt, ähnlich zu vielen natürlichen Systemen, die sich allein aufgrund ihrer eigenen Anpassungsmechanismen entwickeln. Überraschenderweise zeigen natürliche Systeme, die selbst-organisierend sind, oft verschiedene Muster und zeitliche Strukturen, die spontan aus einfachen vordefinierten dynamischen Regeln entstehen. Aktivität scheint sich im Gehirn selbst zu organisieren, da neuronale Aktivitätsmuster und Synapsen viel zu zahlreich und zu variabel scheinen, als dass sie in genetischen Anweisungen hart codiert sein könnten.

Das Gehirn als ein kollektives Phänomen zu untersuchen, ist ein relativ neuer Ansatz in den Neurowissenschaften, und viele Fragen sind noch immer unbeantwortet. Eines der faszinierendsten Merkmale der Mechanismen der Selbstorganisation im Gehirn ist beispielsweise dessen Fähigkeit, das Level der Gesamtaktivität genau zu steuern, sowie das gesamte System unter möglicherweise stark variierender Intensität des Eingangssignals stabil

---

[1] This abstract corresponds approximately to the German translation of Chapter 1.

zu halten. Dieses Merkmal ist sogar noch rätselhafter, da es möglich sein muss innerhalb eines bestimmten Gehirnzustandes verschiedenartige Informationen zu verarbeiten, einschließlich des Lernens und Speicherns externer Inputmuster. Basierend auf der Idee kritischer Phänomene sowie experimenteller Beobachtungen von Anzeichen für Kritikalität ist kürzlich eine Hypothese vorgeschlagen worden, nach der der dynamische Zustand, in dem sich ein gesundes Gehirn befindet, tatsächlich ein kritischer Punkt ist, der sich am Übergang zwischen einer unterkritischen und einer überkritischen Phase befindet. Diese Hypothese ist jedoch immer noch umstritten, da Kritikalität nur indirekt in biologischen neuronalen Netzwerken nachgewiesen werden kann. Darüber hinaus gibt es mehrere experimentelle und theoretische Argumente, die die potenziellen funktionellen Rollen von Kritikalität unterstützen oder unterminieren.

Hier untersuchen wir, wie Anzeichen für Kritikalität aufgrund der durch Plastizität gesteuerten Selbstorganisation in neuronalen Schaltkreisen entstehen, und schlagen neuartige Beziehungen zwischen ihrem Auftreten und der Lern- und Speicherkapazität eines Netzwerks vor. Dazu verwenden wir ein selbst-organisierendes, rekurrentes neuronales Netzwerkmodell, das räumlich-zeitliches Lernen durch die Nutzung biologisch inspirierter Plastizitätsmechanismen ermöglicht. Wir beschreiben die notwendigen Bedingungen für die Aufrechterhaltung dieser Anzeichen in Form experimentell beobachteter lawinenartiger neuronaler Aktivitätsmuster und zeigen, dass sie aus denselben Anpassungsmechanismen resultieren, die das Lernen und das Gedächtnis verbessern. Dies deutet darauf hin, dass der dynamische Zustand neuronaler Schaltkreise an besondere äußere Anforderungen angepasst werden könnte und sollte. Im Zuge des maschinellen Lernens untersuchen wir weitere Anwendungen dieses besonders nützlichen dynamischen Zustands für Sequenz- und Sprachlernen. Abschließend schlagen wir vor, dass die gleiche Kombination von Plastizitätsmechanismen, die für die Gehirnfunktionen auf hoher Ebene verantwortlich ist, eine wesentliche Rolle darin spielt neuronale Schaltkreise hin zu oder weg von einem kritischen Zustand einzustellen.

Anzeichen für Kritikalität resultieren aus plastizitätsgetriebener Selbstorganisation

Im ersten Teil dieser Dissertation wird gezeigt, dass experimentell beobachtbare Anzeichen für Kritikalität, wie etwa nach einem Potenzgesetz verteilte

Aktivitätsbursts, auch „neuronale Lawinen“ genannt, in der Aktivität von selbstorganisierten rekurrenten neuronalen Netzwerken zu finden sind. Durch die Wirkung von synaptischer Plastizität, gehören diese anpassungsfähigen Netzwerke wahrscheinlich zu einer anderen Universalitätsklasse als stochastische Verzweigungsprozesse, die häufig zur Veranschaulichung von neuronalen Schaltkreisen eingesetzt werden. Es wird ausgeführt, dass das verwendete rekurrente neuronale Netzwerk deshalb nicht nur nicht-triviale Aktivitätsmuster aufweist, sondern auch keine Trennung von Input- und internen Zeitskalen besitzt, was die Beobachtung von kleineren Potenzgesetz-Exponenten erklärt. Außerdem wird gezeigt, dass eine Kombination von Hebb'scher und homöostatischer Plastizität verantwortlich dafür ist, das Netzwerk in einen Zustand zu lenken, in dem Anzeichen für Kritikalität erscheinen, diese Kombination jedoch nicht für ihre Aufrechterhaltung im Falle spontaner Aktivität benötigt wird.

Evozierte Aktivität verkompliziert den Sachverhalt. Der Noise-Pegel an neuronalen Membranen bestimmt den dynamischen Zustand des Netzwerks. Dies deutet darauf hin, dass ein Phasenübergang von einem scheinbar superkritischen zu einem rein stochastischem Zustand stattfindet. Weiterhin trägt unstrukturierter Input dazu bei, dass Aktivitätsmuster in einer kurzen Übergangsperiode keine Potenzgesetz-Verteilung mehr aufweisen, was jedoch durch die Wirkung von Plastizität wiederhergestellt wird. Dies verläuft analog zu experimentellen Beobachtungen. Strukturierter Input, wie in einfachen Lernaufgaben, sorgt dafür, dass das Netzwerk in einen Zustand versetzt wird, wo keine Anzeichen für Kritikalität zu finden sind. Diese Ergebnisse des Modells können ähnlich auch experimentell nachgewiesen werden: in vivo Aufnahmen von neuronalen Aktionspotentialen zeigen einen leicht subkritischen Zustand, während Anzeichen für Kritikalität in vitro beobachtet werden. Aus der Perspektive des Gehirns ist so eine Adaption an externe Voraussetzungen von extremer Bedeutung für gesunde dynamische Zustände, da superkritische Zustände mit pathologischen Verhaltensmustern einhergehen. Die Ergebnisse zeigen auf, wie eine Kombination von biologisch inspirierten Plastizitätsreglen diese beiden Phänomene erklären kann, während diese gleichzeitig essentiellen Gehirnfunktionen zugrunde liegen, wie z.B. Lernen und Gedächtnis.

Lernen und Gedächtnis in der Nähe kritischer Dynamik

Die funktionalen Aufgaben kritischer Hirndynamiken sind nach wie vor Gegenstand aktueller Forschung und auch der Schwerpunkt im zweiten Teil dieser Dissertation. Zahlreiche theoretische Studien haben gezeigt, dass die Informationsverarbeitungsleistung in Systemen maximal ist, die in einen kritischen Zustand eingestellt sind. Zu dieser Leistung gehören die Empfindlichkeit für unterschiedliche Eingangssignale, die Informationsübertragung und die Speicherkapazität von Signalmustern. Es ist nicht überraschend, dass es insbesondere in adaptiven Systemen sehr herausfordernd ist, eine Verbindung zwischen Kritikalität, den Anzeichen für Kritikalität und hohen kognitiven Funktionen wie Lernen und Gedächtnis zu ziehen. Auf der theoretischen Seite erfordert Lernen von Aufgaben typischerweise, dass das Inputsignal eine bestimmte Stärke aufweist und die Signalmuster sich nicht einfach von der Dynamik des Modells entkoppeln lassen, was zu getriebenen dynamischen Zuständen führt. Auf der experimentellen Seite verlangt das Messen „neuronaler Lawinen“ lange Aufnahmen neuronaler Aktivität um hinreichend viele Ereignisse zu detektieren, während ein dynamischer Zustand sich schnell durch verschiedene Plastizitätsmechanismen verändern kann. Des Weiteren wird die Sache zusätzlich dadurch verkompliziert, dass kritische Zustände in zwei unterschiedlichen Systemen auftreten, nämlich in selbstorganisierten kritischen Systemen und an Phasenübergängen von geordneten zu chaotischen Dynamiken. Und beide "Definitionen" treten im Allgemeinen, nicht gleichzeitig auf. Aus diesen Gründen werden Verbindungen zwischen Kritikalität und Gehirnfunktionen normalerweise nur indirekt durch Anzeichen für Kritikalität suggeriert und es wurde bisher keine formale Theorie entwickelt.

Hier zeigen wir zunächst, dass das strukturierte Eingangssignal einer zu lernenden Aufgabe, Anzeichen für Kritikaliät, welche normalerweise in spontaner Aktivität von Netzwerken auftreten, zerstört. Somit sind Anzeichen für Kritikalität beim Lernen einer Aufgabe abwesend, was eine interessante Parallele zu in-vivo Aktivität bedeutet. Wichtig ist dabei hervorzuheben, dass die von Plastizität getriebenen Mechanismen der Selbstorganisation die Befähigung des dynamischen Netzwerks zum Lernen von räumlich-zeitlichen Mustern, im Vergleich zu statischen Netzwerken, verbessert. Wir zeigen zusätzlich, dass unser Modell eine verbesserte Gedächtniskapazität aufweist, d.h. dass das Gedächtnis langsamer nachlässt, da das Modell auch temporäre Signalmuster nach langen Verzögerungen wieder aufrufen kann. Dies ist ein

Hinweis darauf, dass die Netzwerkdynamiken nach der Selbstorganisation nahezu kritisch sind, da wir eine logarithmische Skalierung der Gedächtniskapazität als Funktion der Netzwerkgröße beobachtet haben. Solch eine Skalierung ist maximal und es ist bekannt, dass sie nur in rekurrenten Netzwerken oder Reservoir-Systemen auftritt, welche am Übergang von geordneter zu chaotischer Dynamik operieren. Abschließend lassen diese Resultate darauf schließen, dass der dynamische Zustand von neuronalen Schaltkreisen an die Aufgabenanforderungen angepasst werden sollte, um sich in bestimmten Funktionen hervorzutun. Obwohl es nicht erforderlich ist, treten in manchen jener Zustände Anzeichen für Kritikalität auf, da eine kritische Dynamik für die Informationsverarbeitung vorteilhaft ist, während diese scheinbar unkorrelierten Phänomene durch Selbstorganisation aus den selben Plastizitätsmechanismen entstehen können.

Satzbildung und grundlegende Sprachverarbeitung mit selbst-organisierenden rekurrenten Netzwerken

Im letzten Teil dieser Arbeit nutzen wir die kritischen Eigenschaften nachlassender Gedächtniskapazität von rekurrenten neuronalen Netzwerken, die durch Selbstorganisation entstehen, und wenden uns dem maschinellen Lernen zu, indem wir Netzwerke eine einfache Grammatik lernen lassen. Wir zeigen, dass sogar relativ kleine Netzwerke, bestehend aus Hunderten von Neuronen, nicht nur in der Lage sind, künstlich erzeugte Sätze aus einem künstlichen Wörterbuch auf der Ebene von einzelnen Zeichen zu lernen, sondern auch neue, korrekte Kombinationen von Wörtern zu generieren, obwohl es sich um ein deterministisches System handelt. Trotzt der Tatsache, dass dies eine leichte Aufgabe für moderne Deep Learning Ansätze ist, schlagen wir vor, dass biologisch inspirierte Selbstorganisation Einblicke geben kann, wie Lernregeln, die auf der Plastizität des Gehirns basieren, die generelle Architektur solcher rekurrenter Netze verbessern können, und — was vielleicht noch interressanter ist — wie solche Prozesse in sich entwickelnden neuronalen Schaltkreisen ablaufen könnten. Durch letzteres motiviert, erforschen wir eine etwas anspruchsvollere Sprachgenerierungsaufgabe, basierend auf Sprachtranskripten von echten Kindern, und beschreiben wie selbst-organisierende neuronale Netze im Vergleich mit einfachen, zufällig generierten Deep Learning Modellen abschneiden.

Diskussion

In dieser Arbeit zeigen wir, auf welche Art und Weise von Plastizität getriebene Selbstorganisation vielen simultanen Phänomenen zugrunde liegt, wie unter anderen dem Auftreten neuronaler Lawinen, räumlich-zeitlichem Lernen und Verbesserungen des nachlassenden Erinnerungsvermögens neuronaler Netze. Insbesondere zeigen wir, dass in Experimenten häufig beobachtete Anzeichen für Kritikalität in der spontanen Aktivität eines Modells auftreten, das ursprünglich zum Erlernen von Sequence-Learning-Tasks konzipiert war. Diese benötigen allerdings einen bestimmten Noise-Pegel an der neuronalen Membran. Die Anzeichen für Kritikalität hängen nichtsdestotrotz stark vom externen Input ab: während unstrukturierter externer Input sie nur vorübergehend unterbricht, schaltet strukturierter Input sie komplett ab. Diese Erkenntnisse zeigen eine direkte Parallele zu neuronaler Aktivität in-vivo und in-vitro, die auch sowohl Zeichen eines getriebenen subkritischen als auch eines kritischen Zustands aufweist.

Unsere Untersuchung ist motiviert von der Hypothese, dass das Gehirn ein kritisches System ist, welches sich selbst hin zu einem Phasenübergang zweiter Ordnung organisiert. Aber unsere Resultate sind nicht ausreichend um zu beweisen, dass entweder unser Netzwerkmodell oder das Gehirn sich in einem kritischen Zustand befindet, sondern deuten vielmehr darauf hin, dass Selbstorganisation schnell auf den dynamischen Zustand reagiert, der vom externen Input abhängt. Eine solche Anpassung könnte besonders für neuronale Strukturen von Vorteil sein, da sie den verbesserten Informationsfluss im kritischen Bereich ausnutzen könnten, während sie eine stabile Dynamik für eine große Bandbreite an Input-Intensitäten aufrecht erhalten. Wir heben hervor, dass die hier beschriebene Selbstorganisation zu einer logarithmischen Skalierung des nachlassenden Erinnerungsvermögens führt, was bisher nur in Modellen am Rande des Chaos beobachtet wurde. Es ist derzeit unbekannt, in was für einem Zusammenhang dieser Übergangspunkt, der auch kritisch genannt wird, zu Anzeichen für Kritikalität und „neuronalen Lawinen“ steht. Wir legen nahe, dass adaptive Systeme in der Lage sein könnten, beide „Arten“ von Kritikalität miteinander zu vereinen, wenn auch möglicherweise als unterschiedliche dynamische Zustände, und erwarten das zukünftige Arbeit Klarheit über diese formale Beziehung schaffen wird.

Die Charakterisierung von Kritikalität in neuronalen Netzen eröffnet neue Perspektiven in einer Vielzahl von Gebieten. Das Verständnis der Adaption von Netzwerken an Kritikalität kann zu besseren Diagnosewerkzeugen und eventuell Behandlungsmöglichkeiten für Krankheiten führen, die gemeinhin mit nicht-kritischen Dynamiken assoziiert werden, wie z.B. Epilepsie. Zusätzlich entwickeln und verbessern sich viele Gehirnfunktionen, einschließlich des Lernens und des Erinnerungsvermögens, auf der Grundlage derselben Mechanismen, die für die Aufrechterhaltung der Anzeichen für Kritikalität zuständig sind. Eine sorgfältige Beschreibung der notwendigen Bedingungen für deren Auftreten kann zu Erkenntnissen über die Entstehung dieser Funktionen im Gehirn führen und möglicherweise die zugrunde liegenden Mechanismen erklären. Schließlich besitzen auf Kritikalität eingestellte Modelle beeindruckende Fähigkeiten Informationen zu verarbeiten, aber die Forschung über die Selbstorganisation hin zur Kritikalität in neuronalen Netzen steckt noch in den Kinderschuhen. Ein besseres Verständnis ihrer Rolle in Modellen von Gehirn-Strukturen könnte letztendlich zu effizienteren Architekturen für maschinelle Lernverfahren führen. Folglich hoffen wir, dass diese Arbeit den Weg für zukünftige Studien zur Rolle von Kritikalität und Selbstorganisation in unterschiedlichen Arten neuronaler Netze bereitet.

List of Abbreviations

BOLD	**B**lood-**O**xigenated **L**evel **D**ependent activity
EPSC	**E**xcitatory **P**ost-**S**ynaptic **C**urrents
fMRI	**f**unctional **M**agnetic **R**essonance **I**maging
GRU	**G**ated **R**ecurrent **U**nit
iSTDP	**I**nhibitory **S**pike-**T**iming-**D**ependent **P**lasticity
IP	**I**ntrinsic **P**lasticity
IPSP	**I**nhibitory **P**ost-**S**ynaptic **P**otential
LFD	**L**ong-**T**erm **D**epression
LFP	**L**ong-**T**erm **P**otentiation
LSTM	**L**ong **S**hort-**T**erm **M**emory
MC	**M**emory **C**apacity
MCP	**M**c-**C**ulloch & **P**itts
MEA	**M**icro**e**lectrode **A**rray
MLE	**M**aximal **L**ikelihood **E**stimation
MLP	**M**ulti-**L**ayer **P**erceptron
RNN	**R**ecurrent **N**eural **N**etwork
SBP	**S**tochastic **B**ranching **P**rocess
SN	**S**ynaptic **N**ormalization
SOC	**S**elf-**O**rganized **C**riticality
SORN	**S**elf-**O**rganizing **R**ecurrent **N**eural network
SORN$_{\mathbf{L}}$	**S**elf-**O**rganizing **R**ecurrent **N**eural network by Lazar et al. (2009)
SORN$_{\mathbf{Z}}$	**S**elf-**O**rganizing **R**ecurrent **N**eural network by Zheng et al. (2013)
SP	**S**tructural **P**lasticity
STDP	**S**pike-**T**iming-**D**ependent **P**lasticity
VTA	**V**entral **T**egmental **A**rea

Contents

Contributions iii

Acknowledgements v

Abstract vii

Kurze Zusammenfassung ix

Zusammenfassung xi

List of Abbreviations xix

1 A brief introduction 1
- 1.1 Self-organization, criticality, and the brain 1
  - 1.1.1 Research outline 2
- 1.2 Criticality signatures 3
- 1.3 Learn and memory 4
- 1.4 Sentence generation and language processing 5
- 1.5 Discussion 5

2 Self-organization and criticality: from piles of sand to neural circuits 9
- 2.1 Self-organization and phase transitions 10
  - 2.1.1 A very brief introduction to critical phenomena 12
- 2.2 Self-organized criticality 13
  - 2.2.1 The sandpile model 14
  - 2.2.2 Stochastic branching processes 19
  - 2.2.3 Other SOC models and criticality signatures 22
- 2.3 Self-organization in neural circuits 23

2.3.1 Spike-timing-dependent plasticity 24
2.3.2 Homeostasis 27
2.3.3 Structural plasticity 28
2.4 The critical brain hypothesis 29
2.5 Discussion and outline 31

3 Modeling neural activity: self-organizing recurrent neural networks 35
3.1 A brief historical perspective 36
3.1.1 Perceptrons 36
3.1.2 Recurrent neural networks 39
3.1.3 Reservoir Computing 40
3.1.4 Plasticity induced self-organization 41
3.2 The SORN model 42
3.2.1 Network dynamics 42
3.2.2 Plasticity mechanisms 45
3.2.3 External input and spontaneous activity 47
3.2.4 Training the readout layer 49
3.3 Previous studies and model variations 50
3.4 Where do SORNs stand today? 52
3.4.1 Beyond binary networks: leaky integrate-and-fire neurons 53
3.4.2 SORNs vs. deep neural networks 53
3.5 Discussion 55

4 Criticality meets learning: neuronal avalanches in spontaneous and evoked activity 57
4.1 Neuronal avalanches and criticality 58
4.1.1 Power-laws and criticality signatures in experimental data 62
4.1.2 *Ex-vivo* experiments: the role of external input 66
4.1.3 Modeling neuronal avalanches and criticality 68
4.1.4 Maximal information processing 70
4.1.5 Criticality criticisms 75
4.2 Self-organization towards criticality 78
4.2.1 SORNs revisited 79
4.2.2 SORN's spontaneous activity shows power-laws 81

4.2.3 Criticality signatures are not the result of ongoing plasticity . . . 87
4.2.4 Noise level contributes to the maintenance of the power-laws . . . 89
4.3 External input: readaptation and learning . . . 93
4.3.1 Network readaptation after external input onset . . . 93
4.3.2 Absence of criticality signatures under input of learning tasks . . . 95
4.4 More than one critical point? . . . 99
4.5 Discussion . . . 102

5 Learning with plasticity: from random sequences to sentence generation 107
5.1 Criticality in reservoir computing . . . 108
5.1.1 Self-organization in reservoirs . . . 112
5.2 Fading memory capacity . . . 113
5.2.1 Critical capacity scaling with plasticity action . . . 115
5.2.2 A link between neuronal avalanches and memory improvement . . . 117
5.3 Simple grammar learning . . . 119
5.4 SORNs as generative language models . . . 126
5.4.1 Language acquisition . . . 127
5.4.2 SORNs vs. “deep” recurrent neural networks . . . 130
5.5 Discussion . . . 134

6 Conclusion and outlook 139
6.1 Summary . . . 139
6.1.1 Criticality in neural circuits . . . 140
6.1.2 Criticality meets learning and memory . . . 141
6.2 Outlook . . . 142
6.2.1 Criticality and neural circuits . . . 142
6.2.2 Self-organization and computing . . . 144

A Model implementation in python 145
A.1 SORN . . . 146
A.2 SORN-V2 . . . 147

B Fitting power-laws with maximum likelihood estimators 149

C Grammar and language learning 155
C.1 Predefined dictionaries 155
C.2 The CHILDES dataset 157
C.3 A simple GRU network 158

Chapter 1

A brief introduction

> If in physics there's something you don't understand, you can always hide behind the uncharted depths of nature. You can always blame God. You didn't make it so complex yourself. But if your program doesn't work, there is no one to hide behind. You cannot hide behind an obstinate nature. If it doesn't work, you've messed up.
>
> Edsger W. Dijkstra

1.1 Self-organization, criticality, and the brain

Understanding how the brain works is one of the most challenging problems in modern science. As a complex system composed of hundreds of billions of non-linear information processing units, connected by hundreds of trillions of synapses, the brain's collective dynamics resemble many classic systems from statistical physics. In addition, this enormous system is highly dynamical and constantly adapts itself revealing a large repertoire of spatio-temporal activity patterns, which are the basis for adaptive behaviour. Multiple adaptation mechanisms are known to act simultaneously and regulate synapses and overall brain activity, underlying virtually all vital functions, from basic muscle movement to high level cognitive processes. These emergent neural phenomena take place independently of external control, similarly to many other widely studied natural systems which evolve solely due to their own adaptive mechanisms. Surprisingly, natural self-organization phenomena typically show various patterns and temporal structures, frequently with apparent purpose, that arise spontaneously from simple predefined dynamical rules. This is seemingly what occurs in the brain, as neuronal firing patterns and synapses are far too numerous and variable to be hard coded into genetic instructions.

Studying the brain as a collective phenomenon is a relatively new approach in theoretical neuroscience, and many questions still remain to be answered. For instance, one of the most intriguing features of the brain's self-organizational mechanisms is their ability to precisely control the overall activity level and keep the whole system stable under a wide range of different input intensities. This feature is even more puzzling given that any particular operational state must maintain, or even improve, various information processing capacities, including learning and storing external input patterns. Based on the notion of critical phenomena and experimental observations of criticality signatures, a recent hypothesis proposed that the dynamical state in which a healthy brain operates is, in fact, a critical point, poised at the transition between a subcritical and a supercritical phase. However, this *critical brain hypothesis* is still controversial since criticality can only, if at all, be detected indirectly in biological neural circuits, and multiple experimental and theoretical arguments have emerged to support or undermine its potential functional roles.

1.1.1 Research outline

In this thesis, we investigate how criticality signatures emerge due to plasticity driven self-organization in neural circuits and propose novel links between their occurrence and a network's learning and memory capacities. To do so, we employ a self-organizing recurrent neural network model which is capable of spatio-temporal learning by taking advantage of biologically inspired plasticity mechanisms. We describe the necessary conditions for the maintenance of these signatures, in the form of experimentally observed neuronal avalanches, and show that they result from the same adaptation mechanisms that improve learning and memory, what suggests that the dynamical state of neural circuits could and should be adapted to particular external requirements. In a turn towards machine learning, we further explore applications of this particularly useful dynamical state for sequence and language learning. Finally, we conclude by proposing that the same combination of plasticity mechanisms responsible for high level brain functions plays an essential role in tuning neural circuits towards and also away from a critical state.

1.2 Criticality signatures arise from plasticity driven self-organization

In the first part of this thesis (Del Papa et al., 2017), we show that experimentally observed criticality signatures, power-law distributed bursts of activity, or neuronal avalanches, occur in the activity of self-organizing recurrent neural networks. Due to the action of synaptic and intrinsic plasticity, these adaptive networks likely belong to a universality class distinct from stochastic branching processes, which are a common toy model used in analogy to neural circuits. As a consequence, our model not only displays non-trivial dynamics but also lacks any separation of time scales between input and internal dynamics, which explains the smaller power-law exponents we observe. Nonetheless, we show that a combination of Hebbian and homeostatic plasticity is responsible for driving the network towards a state in which criticality signatures appear, but is not required for their maintenance in the case of spontaneous activity.

The case of evoked activity tells a more complicated story. First, the level of neuronal membrane noise controls the network's dynamical state, suggesting that a phase transition from a seemingly supercritical to a purely stochastic state takes place. Second, unstructured input breaks down the power-laws during a short transient period, but plasticity quickly brings them back, mimicking experiments. Third, structured input of simple learning tasks is enough to drive the model to a distinct regime where no signs of criticality appear. These findings, interestingly, have experimental correlates: *in-vivo* spike recordings indeed show a driven, slightly subcritical regime, while criticality signatures are observed in *in-vitro* setups. From the brain's perspective, such adaptation to external requirements could be extremely important for healthy dynamics, as signs of supercriticality are known to occur during epileptic regimes. Our results suggest how a combination of biologically inspired plasticity rules is able to account for both phenomena while also underlying essential brain functions, including learning and memory.

1.3 Learning and memory near critical dynamics

The functional roles of critical brain dynamics are still subject of ongoing research, and also our focus in the second part of this thesis (Del Papa et al., 2017, 2019). Multiple theoretical studies have shown that information processing capacities are maximal in systems tuned to criticality, including their sensitivity to various inputs, information transmission, and pattern storage capacity. Not surprisingly, linking criticality and its signatures to high level cognitive functions such as learning and memory is much more challenging, especially in adaptive systems. On the theoretical side, learning tasks typically require particular input levels and patterns that are not easily disentangled from a model's dynamics, resulting in driven dynamical states. On the experimental side, the measurement of neuronal avalanches demand long recordings of neural activity in order to detect a sufficient number of events, while a dynamical state can be quickly modified by various plasticity mechanisms. To make matters more complicated, critical points commonly appear in two different types of systems — self-organized criticality phenomena and phase transitions between ordered and chaotic dynamics — and both "definitions" do not co-occur in general. For these reasons, links between criticality and brain functions are usually suggested only by indirect indications of criticality, and no formal theory has been developed so far.

Here, we first show that the structured input of learning tasks breaks down the criticality signatures otherwise found in a network's spontaneous activity. Thus, when performing learning tasks, signatures of criticality are absent, reinforcing the interesting parallel with *in-vivo* activity. Importantly, the same plasticity driven self-organizing mechanisms improve spatio-temporal learning when compared to static networks. We additionally show that our model, when poised at a phase transition state where neuronal avalanches occur, has an improved fading memory capacity and is able to recall temporal inputs after longer delays. This points to near critical dynamics after self-organization, since we also observed a logarithmic scaling of the memory capacity as a function of the network size. Such scaling is maximal and is known to occur only for recurrent networks or reservoirs operating at the transition between ordered and chaotic dynamics, but may also emerge after plasticity actions. Finally, these results imply that the dynamical state of brain circuits or neural networks should be adapted to task requirements in order to excel at particular functions. Although not a requirement, criticality signatures may appear in some of those states, since critical dynamics is beneficial for information processing, while these seemingly uncorrelated phenomena can result from self-organization due to the same plasticity mechanisms.

1.4 Sentence generation and basic language processing with self-organizing recurrent neural networks

In the last part of this work, we take advantage of the critical-like fading memory scaling of recurrent neural networks after self-organization and turn towards machine learning, by exploring their applications for simple grammar learning tasks. We show that even relatively small networks of hundreds of neurons are not only able to learn artificially created dictionaries of sentences at the character level, but also generate new, correct combinations of words, despite being deterministic systems. Although this is an easy task for modern deep learning frameworks, we suggest that biologically inspired self-organization might give insights on how learning rules based on brain plasticity can improve general recurrent network architectures and, perhaps more interestingly, on how this process might take place in developing neural circuits. Motivated by the latter, we also explore a more challenging language generation task based on speech transcripts from real infant -directed language, and describe how self-organizing neural networks perform compared to simple deep learning models.

1.5 Discussion

In this work, we show how self-organization due to biologically inspired plasticity underlies a combination of phenomena, including the occurrence of neuronal avalanches, spatio-temporal learning, and improvements in the fading memory of recurrent neural networks. In particular, we show that the most common experimentally measurable criticality signatures also occur in the spontaneous activity of a model that was initially conceived for sequence learning tasks, but require a specific neuronal membrane noise level to be maintained. These signatures, nonetheless, depend heavily on the external drive: while unstructured external input only transiently breaks them down, structured input of simple learning tasks abolish them completely. These findings have a direct parallel with spiking activity *in-vivo* and *in-vitro*, which show signs of a driven subcritical and a critical state, respectively.

Our investigation is partially motivated by the critical brain hypothesis, which states that the brain is a critical system that self-organizes to a second-order phase transition point, poised between subcritical and supercritical dynamics. However, our results are insufficient to prove that either our network model or the brain is indeed at a critical state, and instead suggest that self-organization quickly acts on the dynamical state depending on the external input condition. Such adaptation could be particularly beneficial for neural circuits, as they could take advantage of the improved information processing abilities of criticality while maintaining healthy and stable dynamics for a range of input intensities and patterns. Importantly, we highlight that self-organization as described here results in a logarithmic scaling of the fading memory capacity, which was previously observed only for models tuned to a transition point at the edge-of-chaos. It is currently unknown how this transition point, also referred to as critical, generally relates to criticality signatures in the form of neuronal avalanches. We suggest that adaptive systems might be able to combine both "types" of criticality, but possibly as different dynamical states, and expect that future work will clarify the formal relationship between these phenomena.

The characterization of criticality in neural circuits opens new perspectives in many areas. Understanding how networks adapt towards criticality might lead to better diagnostic tools and eventual treatments for diseases whose symptoms are associated with deviations from a critical dynamics, such as epilepsy. In addition, multiple brain functions, including learning and memory, seem to emerge and improve based on same mechanisms responsible for self-organization and maintenance of criticality signatures. A careful description of the necessary conditions for their occurrence can lead to insights into the emergence of these functions in the brain at the network level, and potentially shed light on their underlying mechanisms. Finally, models tuned to criticality have repeatedly proved to possess powerful information processing abilities, but the study of self-organization towards criticality in neural networks is still in its infancy. We believe that a better understanding of its role in simple models of brain circuits might eventually lead to new, more efficient architectures in subareas of machine learning. Thus, we hope this work will pave the way for further studies on the roles of criticality and self-organization in various types of neural networks.

Chapter 2

Self-organization and criticality: from piles of sand to neural circuits

> In such a product of nature every part not only exists by means of the other parts, but is thought as existing for the sake of the others and the whole, that is as an (organic) instrument. Thus, however, it might be an artificial instrument, and so might be represented only as a purpose that is possible in general; but also its parts are all organs reciprocally producing each other. This can never be the case with artificial instruments, but only with nature which supplies all the material for instruments (even for those of art). Only a product of such a kind can be called a natural purpose, and this because it is an *organized* and *self-organizing* being.
>
> Immanuel Kant

Patterns can be observed in virtually all natural systems composed of multiple entities. In fact, the attempted description of emergent order in nature is arguably the inception of many modern scientific areas of investigation, including both physics and neuroscience. Rather surprisingly, spatio and temporal structures sometimes occur spontaneously and seemingly with purpose, as they result in essential functions or properties for a given collective system. In the natural sciences, such broad process has been named *self-organization* (Haken, 2008), since it develops in an unguided and unsupervised manner, and many studies have focused on finding general principles that govern all different observed self-organizing phenomena. After all, if generic principles of self-organization exist and can be applied to varied distinct systems, independently of their structure or particular dynamics, the path towards a unifying theory of everything could be at the horizon. In reality, however, general self-organization rules have been found only for specific classes of systems, but have undoubtedly helped their development towards new directions. In particular, one of the most debated occurrences of self-organization comes from neuroscience: do complex neural systems obey general rules or do particular individual mechanisms dictate their behavior? Do distinct "neural" self-organizational dynamics exist? And if so, do they always tend to converge towards some special states? In this chapter, we provide an overview of self-organization in nature and physical systems, going from toy models to biological neural circuits, in order to lay the foundations for further investigation of self-organizing mechanisms in neural networks. We discuss the most popular hypothesis on large scale self-organization in the brain: the *critical brain hypothesis*. The hypothesis, which has been first proposed relatively recently by Beggs and Plenz (2003), has become the focus of many studies and shed light on possible desirable dynamical states in which the brain could, in principle, operate. We finally discuss potential implications of these self-organization mechanisms, bringing forward the motivation for our study of criticality signatures in biologically inspired neural networks.

2.1 Self-organization and phase transitions in nature

From the curious geometric patterns in snow crystals to the self-dynamics of formation of public opinion, studies have shown that complex systems self-organize following various rules, depending on their boundaries, initial conditions, and parameter values (Haken, 2008). These systems' behaviors typically emerge as a collective function of simpler, mostly local interaction processes without any external guidance[1]. The result of a self-organization process can be illustrated, for example, by complex spatial patterns observed

---

[1] More precisely, this is a related process called *emergence*, and although similar, it is not equivalent to self-organization, as systems can self-organize due to internal mechanisms into states in which no collective property emerges (Crommelinck et al., 2006).

in lizard skins, beach sand, or even in brain activity (Fig. 2.1).

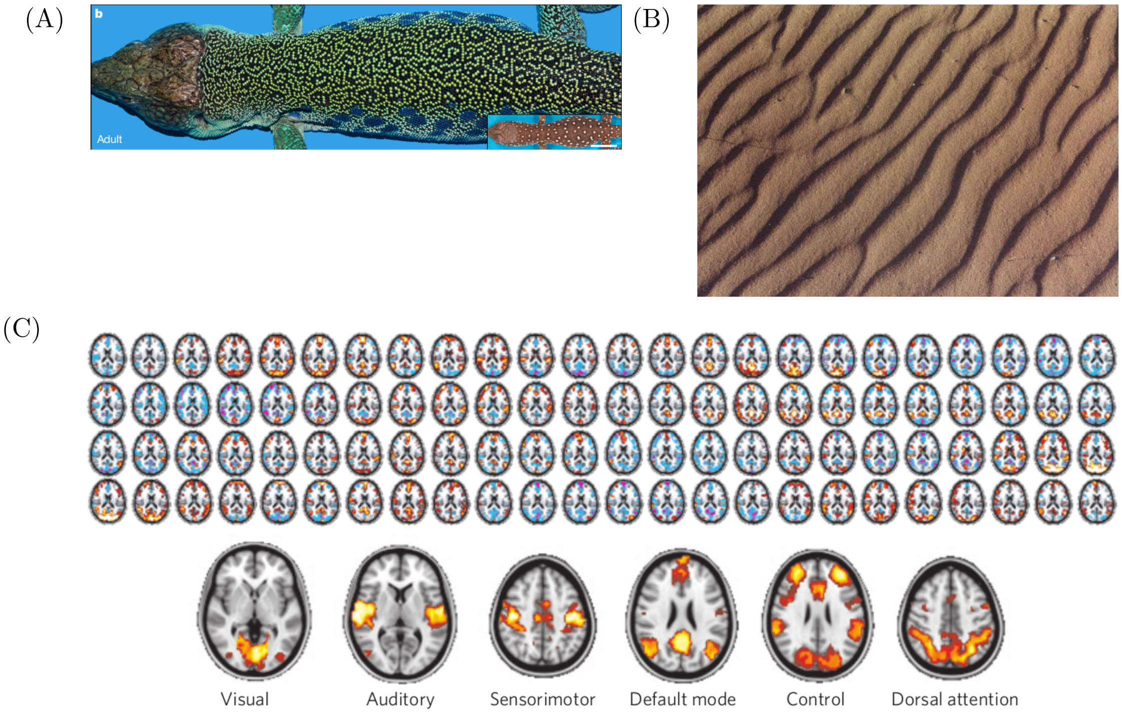


Figure 2.1: **Examples of self-organization and emergence in diverse natural systems.** (A) Color pattern emergence in adult ocellated lizards (inset shows a juvenile lizard, for comparison). The emergence occurs mostly due to local interaction rules among skin cells, resembling a cellular automaton. Figure reproduced from Manukyan et al. (2017) with permission. (B) Sand dune ripples formed by wind action in a desert, with wavelengths typically between 10 and 15 cm and heights of a few mm. Figure reproduced with permission from Yizhaq et al. (2004). (C) Large-scale emergent brain networks. The top sequence of images shows increases (red) and decreases (blue) compared to the mean fMRI blood-oxygenation level dependent (BOLD) activity during consecutive resting brain recordings (2.5 seconds per image). Bottom images are the results of linear correlations between the activity of small regions within the networks of interest and the rest of the brain, corresponding to six main systems: visual, auditory, sensorimotor, default mode, executive control and dorsal attention. Figure reproduced from Chialvo (2010) with permission.

The development of complex patterns requires, naturally, a minimal set of interaction rules among the system's components, which is possibly the main reason why computational toy models have become popular to explain and reproduce self-organization. Generally, local self-enhancement (i.e., positive feedback) and long-range inhibition (i.e., competition for limited resources) have been proposed to be driving forces behind the generation of organized regions and boundaries in many biological systems (Meinhardt, 2008). Some models, additionally, require a specific set of parameters to reproduce particular self-organization patterns and have distinct structure and properties, or *phases*, depending on their tuning. The transition between phases, which modifies the system's collective behavior, can happen quite suddenly and be the result of critical phenomena.

2.1.1 A very brief introduction to critical phenomena

Critical phenomena, although not exclusively, are commonly the result of second order phase transitions in equilibrium, when the second derivative of an order parameter is discontinuous (in contrast to first order phase transitions, in which the first derivative is discontinuous). At critical points, quantities such as the correlation length between units are known to diverge, while the overall dynamics slows down and scale-free properties are known to appear (Scheffer et al., 2009). Perhaps the most famous and seminal example of such transitions is the Ising model (for ferromagnetic-paramagnetic transitions). This model, which by no means we aim to explain here in mathematical detail, describes the collective behavior of classical two-state spins, and how their local neighbor interactions result in different dynamics, or phases, as a function of the temperature (which serves as a control parameter). Regarding the system state in terms of "up" and "down" spin states, each phase follows a characteristic pattern (see Fig. 2.2), with unique dynamics at a critical point. First, at high temperatures, individual spin states change continuously, almost at random, resulting in null mean magnetization. Second, at low temperatures, the system is ordered and exhibits large regions of spins with the same orientation, resulting in magnetization with strong stability over time. Last, at an intermediate critical temperature, the system exhibits distinct spatio-temporal patterns with characteristic properties, such as scale invariant fluctuations of its magnetization and power-law distributed correlations of spins. Additionally, at this critical point, the spin system has the highest susceptibility and a single perturbation can, with a given probability, be propagated and reshape the entire system switching its magnetization direction.

These dynamical properties have analogies in virtually all critical phenomena, both self-organized or externally driven. In particular, one can easily see that properties such as maximum correlation length can be qualitatively associated with improvements in information transmission and that such analogies could be also valid in neural systems. In fact, this is the origin of the critical brain hypothesis. Nonetheless, before explaining how this hypothesis was formulated and how self-organization occurs in the brain, we first describe a particular class of toy models that maintain a critical point without external tuning, and that have drawn a significant amount of attention since their proposal.

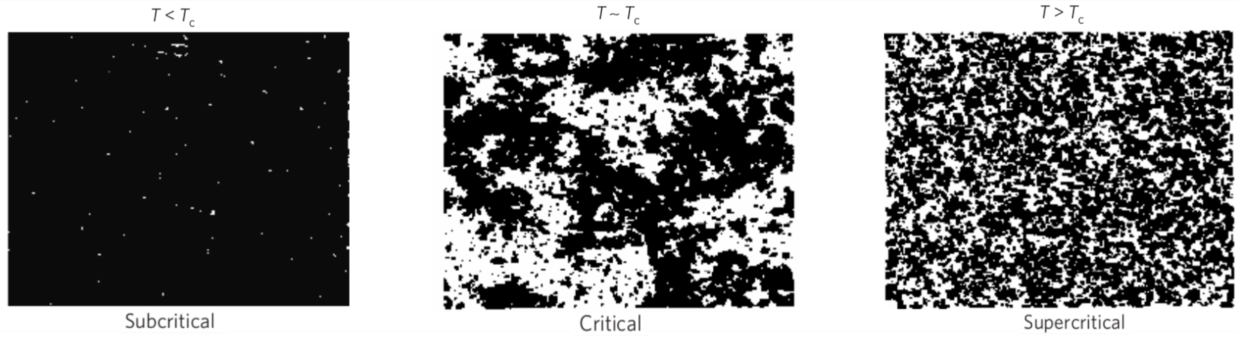


Figure 2.2: **Three phases of the 2D Ising model.** Snapshots of two-state spin configurations for three different temperatures (subcritical, critical, and supercritical). Subcritical and supercritical temperatures result in relatively homogeneous states, while a system at critical temperature T_c displays heterogeneous regions. Figure reproduced with permission from Chialvo (2007).

2.2 Self-organized criticality

As discussed previously, some natural systems spontaneously self-organize into states that may exhibit various spatio-temporal patterns. Under certain conditions, these self-organization states, being at equilibrium, might exhibit similar properties to systems at second order phase transition points, thus resulting in critical phenomena. Such special class of systems has been first described by Bak et al. (1987) and the phenomenon has been named Self-Organized Criticality (SOC), as an attempt to provide unifying principles guiding the behavior of collective systems. Although such attempt resulted in rather bold and incorrect claims about a new universal theory of complex behavior (Frigg, 2003), SOC has since become a description of a group of models that relate via formal analogy due to their internal mechanisms.

Importantly, the universal behavior observed in SOC systems is the result of their diverging correlation lengths (Marković and Gros, 2014). Since the correlation length for those systems is larger than any microscopic local interactions, a collective behavior with scale-free properties appears. In fact, as we will demonstrate with examples, SOC systems are not typically at an equilibrium point, since they exchange energy (or information) continuously with the environment (but, rather remarkably, they are conservative systems — see Bonachela et al. (2010)), which makes them a distinct class of models different from systems displaying “classic” critical points.

When at non-equilibrium states, SOC systems tend to not be analytically solvable, and numerical solutions are often considered on a case by cases basis. However, those systems fall under the same universality notion in statistical physics: systems whose large-scale properties are independent of their particular dynamical details when near a critical point (Kadanoff, 1990). Thus, SOC systems share both scaling functions and their critical exponents, i.e., their exponents describing physical measurable quantities as order parameters. An immediate conclusion of such property is that simple toy models can be used to infer many aspects of the critical behavior of real complex systems, given that they share the same universality class. Therefore, we continue our discussion by describing the model that introduced the SOC nomenclature, the Bak–Tang–Wiesenfeld sandpile model.

2.2.1 The sandpile model

The classic case of SOC systems is the Bak–Tang–Wiesenfeld sandpile model (Bak et al., 1987), also called Abelian Sandpile Model. This toy model exemplifies a general class of self-organizing systems that evolve into a critical point by simulating the growth of theoretical piles of sand on a finite grid, as in a cellular automaton. Intuitively, the model starts from an empty two-dimensional grid, on which unitary grains of sand are randomly and consecutively dropped. When any particular grid cell has many more grains than its nearest neighbors, the column topples, and its sand is distributed equally among neighboring cells. Eventually, the toppling of one column induces the toppling of others in a chain reaction and big avalanche events that span the entire system occur. By definition, a grain of sand that reaches the border of the grid "falls" off the system, countering the addition of new grains and resulting in the conservation of sand (on average). This state is defined as SOC: self-organized in the sense that it occurs spontaneously, without external fine tuning, and critical in the sense that local events (the drop of a single grain of sand) may be propagated and affect the entire system through avalanches, similar to perturbations in the Ising model. Note that although the self-organizing nomenclature implies independence of external influence, the addition of new grains, in fact, mimics the external drive[2], a result of the interaction with the surrounding environment.

Formally, the sandpile model can be generalized to d dimensions. Considering a grid of size L, the height $z(\mathbf{r})$ of a cell at position $\mathbf{r} = (r_1, r_2, .., r_d)$, $r_i < L$, $0 < i \leq L$, is updated according to the rule:

$$z(\mathbf{r}, t+1) = z(\mathbf{r}, t) + \delta z \tag{2.1}$$

in which δz is the number of sand grains/energy added to the model at each discrete update step. After the addition of each new grain, two scenarios may occur. First, if the cell in which the grain was added satisfies the condition $z(t) < z_{\mathrm{T}}$ for a fixed threshold z_{T}, nothing happens and the dynamics continue, adding the next grain to the system. Second, if the height of the cell exceeds its threshold, an avalanche event starts, and toppling occurs with the following update rules:

$$z(\mathbf{r}, t) \leftarrow z(\mathbf{r}, t) - z_{\mathrm{H}} \tag{2.2}$$

$$z(\mathbf{r} + \mathbf{e}, t) \leftarrow z(\mathbf{r} + \mathbf{e}, t) + \beta z_{\mathrm{H}} \tag{2.3}$$

in which z_{H} is the number of grains that topple to the neighbor cells ($|e| = 1$ is a unitary vector, depending on the model's topology in the most general case) and β is a transmission constant. After the first toppling, any cell that also reaches its threshold also topples, thus creating the chain reaction. The update step is over when $z(\mathbf{r}, t) < z_{\mathrm{T}}$ for all $\mathbf{r}$, and the model dynamics continues to the next step. Although the cell in which each grain of sand drops can be chosen at random, depending on the variation of the model,

---

[2] The number of sand grains per cell is, roughly speaking, its local energy or stress level.

if $z_{\mathrm{H}} = z_{\mathrm{T}}$ the toppling process is deterministic and the particular order of toppling does not affect the final system configuration (thus the Abelian model nomenclature).

Even though the sandpile model can be constructed with any given topology, for the sake of brevity we only reproduce here the two-dimensional square grid case. The study of sandpile structures is an interesting field of its own, and many more realistic complex systems, including Erdős-Rényi random graphs and small-world networks, have been shown to belong to the same universality class as high-dimensional lattices in the thermodynamic limit ($L \to \infty$). Different toppling rules, however, can lead to different critical exponents, see Marković and Gros (2014) and references within for an overview and more details on the sandpile dynamics in those systems. Back to the two dimensional case ($d = 2$), setting $\beta = 1/2d$ assures local conservation of grains/energy, although energy dissipation still occurs at the border (in practice, by setting $z(\mathbf{r}) = 0$ if $\mathbf{r}$ belongs to the system's borders). We can set $z_{\mathrm{T}} = 4$ (for convenience, in practice bigger values also result in SOC (Frigg, 2003)) and measure the scaling of observable quantities in numerical simulations, such as the avalanche size S (number of topplings during one avalanche — Fig. 2.3A) and the avalanche duration T (number of time steps until stability is reached — Fig. 2.3B). The scaling of these observables reveals important properties of the sandpile model and its universality class. First, a power-law scaling means scale free behavior, i.e., the scaling of avalanche sizes and durations is independent of the system size, although their exponents depend on the particular parameters of the model. Second, the average avalanche size $\langle S \rangle$ also follows a power-law (with positive exponent) as a function of the avalanche duration (Fig. 2.3C), which determines the critical exponents of the model. In fact, this scaling is a consequence of the finite size scaling of the model, which should hold independently of the time scale (Sethna et al., 2001). Considering a power-law scaling with exponents τ for the avalanche size and α for the avalanche duration, the relation $\frac{d}{dT}\langle S \rangle(T) = a\frac{\langle S \rangle(T)}{T}$ should hold for some constant a. This yields $\langle S \rangle(T) = S_0 T^a$, where $a = \frac{\alpha-1}{\tau-1}$ is a critical exponent of systems of the same universality class as the sandpile model[3]. Third, a relatively small numerical simulation (on a square lattice

---

[3] In fact, this critical exponent is related to other critical exponents of the model via renormalization theory and is commonly written as their multiplicative combination $\frac{\alpha-1}{\tau-1} = \frac{1}{\sigma \nu z}$. This derivation is, however, beyond our scope in this brief introduction, and a formal derivation of the critical exponent values and their relations can be found in Sethna et al.

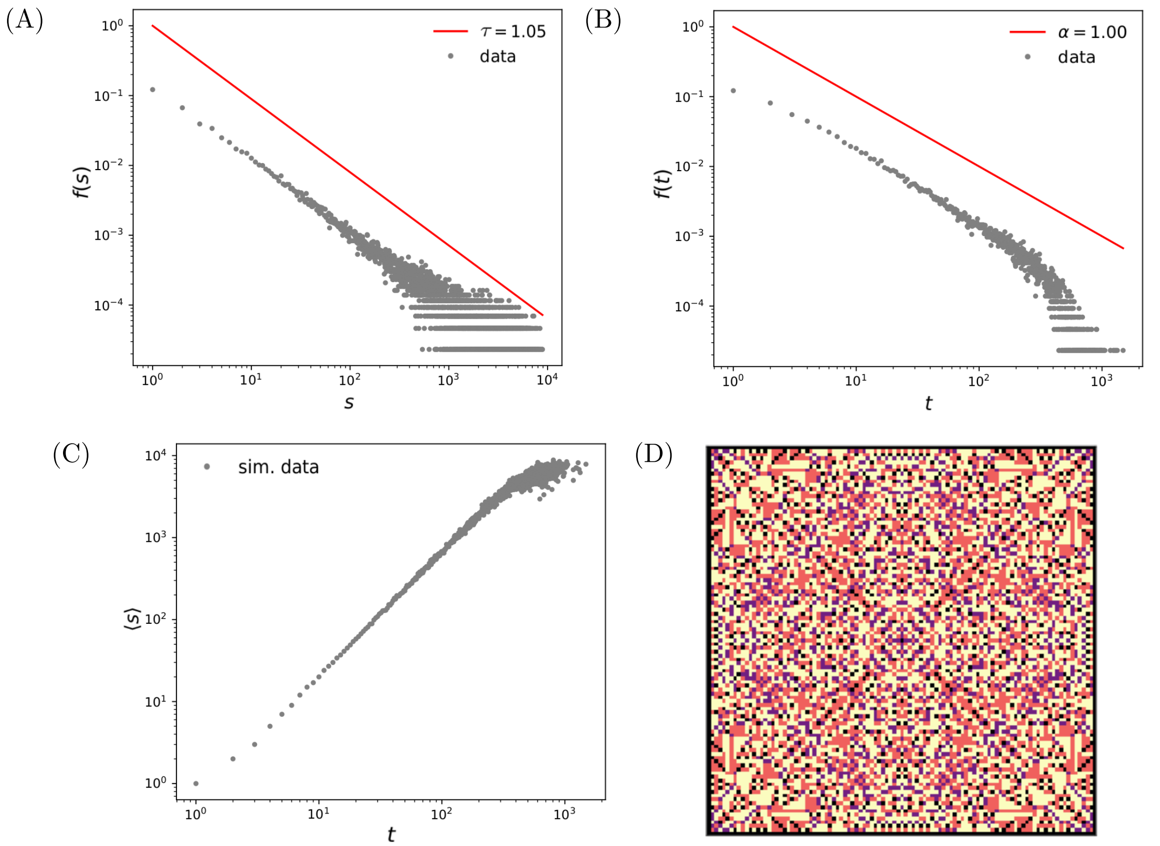


Figure 2.3: **Scaling in the Bak–Tang–Wiesenfeld sandpile model.** (A) Normalized avalanche size distribution for a 100×100 square lattice, after 10^5 grains dropped on random cells. Gray points show raw numerical results and α shows the fitted power-law exponent (via maximum likelihood estimators — see Appendix B). (B) Normalized duration distribution for the same grid. (C) Power-law scaling of average avalanche size $\langle S \rangle$ as a function of avalanche duration, for the raw simulation data. (D) Snapshot of emergent symmetrical patterns when grains are continuously dropped in the middle of the lattice. Colors indicate the number of grains/energy in each cell. Simulation code can be found at `https://github.com/delpapa/sandpilemodel`.

of 100 × 100 cells) illustrates the distributions' tail effect in finite systems. Although the avalanche size and duration follow a power-law for a number of orders of magnitude, their tails reveal biases due to finite size cut-offs and rounding effects, requiring special fitting algorithms to avoid propagating those biases when estimating their exponents (see Appendix B). Of course, bigger systems yield power-laws that span more orders of magnitude, and infinite systems display no cut-off effects. Fourth and last, depending on the initial conditions and lattice shape/size, the sandpile model can result in interesting symmetrical patterns (Fig. 2.3D) or fractal structure if enough grains are dropped on the system.

Importantly, the dynamics of the sandpile model described so far relies on an essential assumption: the separation of time scales between external drive (the dropped grains of sand) and internal dynamics (the topplings and avalanches). This separation means that no grains are dropped, or equivalently no energy is added to the system, while its own dissipation process is still ongoing, creating pauses between each time step. This property, which is rather unique to toy models and absent from many real complex systems such as living neural circuits, is essential for keeping the model at a critical point. In fact, considering the density of active states (states in which avalanches only stop due to the dissipation at the borders) as an order parameter, it is possible to show that a second order phase transition appears and can be tuned via a mechanism that balances external drive and dissipation (Marković and Gros, 2014). Thus, while a system in which external drive and dissipation occur separately at a phase transition point (i.e., they are infinitely separated by construction) can always be tuned to criticality, its scaling properties and SOC state cannot be, in principle, generalized to other complex systems without separation of time scales.

Finally, it is important to mention that although sandpile models were inspired by sand patterns and avalanches, they are toy models and SOC has received much criticism regarding being a ubiquitous theory of nature (Frigg, 2003). For instance, even real piles of sand only display SOC and the aforementioned scaling properties up to a certain size limit, and virtually every property of the toy model breaks down in the limit of large real piles of sand (Held et al., 1990). Additionally, piles of rice only show apparent SOC and power-law distributions of avalanche sizes in the limit of elongated rice grains, but not for round ones (Frette et al., 1996), suggesting that SOC is

---

(2001); Marković and Gros (2014) and references within.

indeed dependent on the details of the system, while deviations from critical points cannot be disregarded as experimental noise or imprecision. Instead, those differences are rooted in the sandpile model simplicity: it does not account for physical interactions between grains or the limited speed with which they move or accelerate. Thus, even though we can conclude that theoretical piles of various grains display SOC and power-law scaling, the same cannot be stated for experimental ones.

2.2.2 Stochastic branching processes

We continue by describing another important class of models that displays critical behavior and has perhaps a more straightforward analogy to neural activity: branching processes (Harris, 2002; Beggs and Plenz, 2003). Formally, a branching process is defined as a multiplicative Markov chain of positive random integer values, in which a generation of active units (or particles or neurons) at a given time step yields a new generation of active units at the immediate posterior time step (Marković and Gros, 2014). More specifically, a stochastic branching process (SBP), as studied by Haldeman and Beggs (2005), describes the mapping of branching processes onto a probabilistic network of simple neurons, in which each neuron i, when active, has a fixed probability p_{ij} of activating its neighbor neuron j at the subsequent time step. The system's order is measured via its branching parameter σ, given by:

$$\sigma_t = \frac{n_t}{n_{t-1}} \tag{2.4}$$

in which n_t is the number of active neurons at time step t. The system is considered to be critical when it maintains its activity level over time, i.e., when on average $\sigma \approx 1$. Likewise, the system is subcritical when, on average, activity dies out ($\sigma < 1$) and supercritical when activity eventually takes over all neurons ($\sigma > 1$). For a network of N neurons, the branching parameter can be estimated for each single neuron i by considering the sum of its connectivity probabilities, $\sigma_i = \sum_{i=1}^{N} p_{ij}$. More formally, this control parameter is the extinction probability, i.e., the probability of activity reaching zero in the limit of infinity time steps. From this probability one may estimate the Lyapunov exponents of the system, from which it can be shown that as long as the condition $\sum_{i=0}^{N} k_i p_{ij} = 1$ is satisfied for all neurons j, the system is critical (i.e., its Lyapunov exponents are zero) and power-law distributions with exponents $\alpha_S = -1.5$ for avalanche sizes and $\alpha_T = -2$ for avalanche durations appear in infinitely large systems (see Otter (1949) and Marković and Gros (2014) for the derivation). For simplicity, following the model by Priesemann et al. (2014), we can assume that all neurons have the same fixed number of neighbors k, which can be randomly chosen at each time step and activated with the same fixed probability p. This results in a control parameter $\alpha = p \times k$, which is also unitary at the critical point. Similarly to the sandpile model, branching processes yield power-law distributions with finite size cut-offs and a characteristic critical exponent when at criticality, which can be approximated by simulations of relatively small systems (Fig. 2.4: $N = 2500$, $k = 4$, and 10^6 time steps). Remarkably, even a relatively small and easy to simulate network is already capable of approximating the theoretical critical exponents, both for the avalanche sizes (Fig. 2.4A) and durations (Fig. 2.4B). Small deviations from criticality already result in different distributions, for which power-laws are not the best fit. For the subcritical case, the decay is faster than a power-law, resembling an exponential distribution. For the supercritical case, bigger avalanches appear much more frequently, and, mathematically speaking, there is a non-zero probability of infinite duration.

Branching processes are particularly insightful in the modeling of criticality in neural circuits for mostly three reasons. First, they have a straightforward analogy to biological neural networks, with neurons and transmission probabilities that can easily be seen as simplified synapses. Second, the simplicity of the model allows for a good understanding of the role each parameter plays in the self-organization process and how they affect the tuning towards a critical point. Critical exponents can be identified and measured based uniquely on the size and duration of experimental events. Last, it is relatively easy to include interactions with the external environment, by adding a probability of dissipation and an external drive parameter (see the model by Priesemann et al. (2014) for an analysis of these additional parameters). Thus, the analogy of neural circuits as branching processes lies at the inception of the critical brain hypothesis (Beggs and Plenz, 2003) and they have become an important tool to study self-organization towards criticality in the brain.

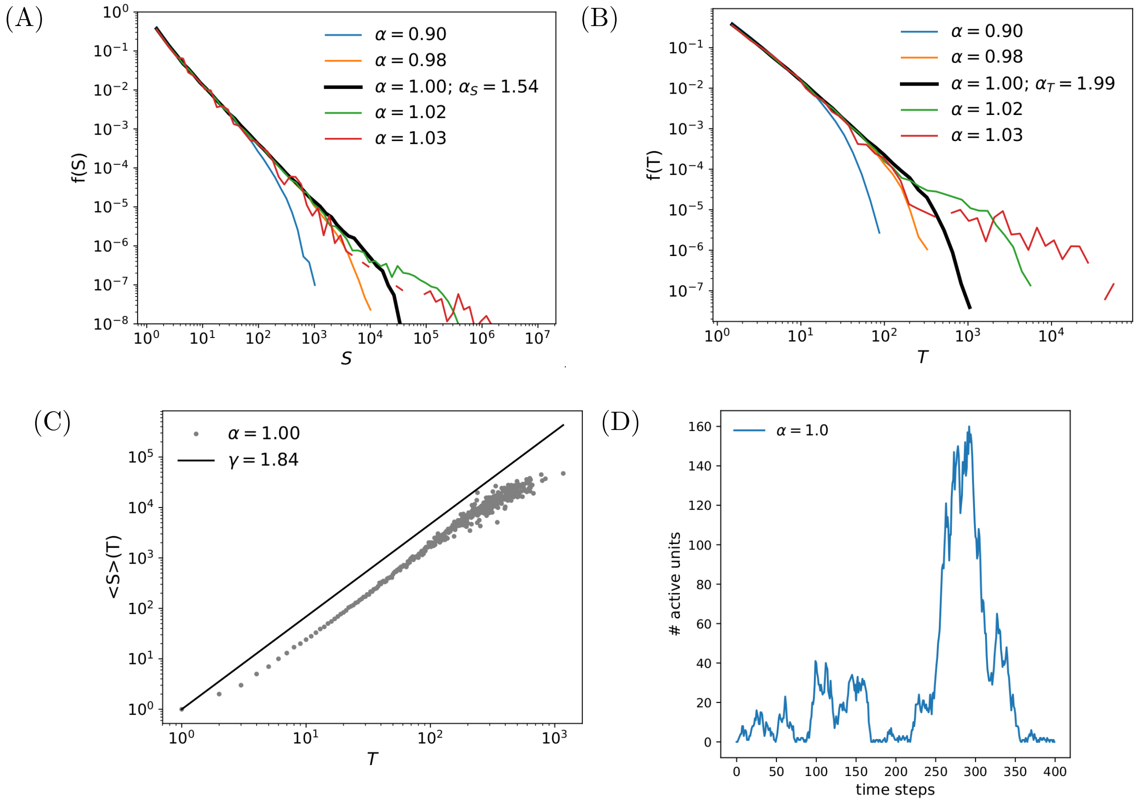


Figure 2.4: **Avalanches in stochastic branching processes.** (A) Normalized avalanche size distribution for a system at criticality ($\alpha = 1$, power-law exponent $\alpha_\mathrm{S} \approx 1.54$) and at slightly subcritical ($\alpha = 0.98$ and $\alpha = 0.90$) or supercritical ($\alpha = 1.02$ and $\alpha = 1.03$) regimes. (B) Analogue distributions for avalanche duration (power-law exponent $\alpha_\mathrm{T} \approx 1.99$). (C) Scaling of the average avalanche size $\langle S \rangle$ as a function of duration (power-law with positive exponent $\gamma \approx 1.84$). (D) Sample activity distribution (number of active units) for a system at criticality, where each avalanche is initiated immediately after the previous one is over (i.e., one time step after the activity reaches zero). Plots show stochastic branching processes with the following parameters: $N = 2500$, $k = 4$, and 10^6 time steps.

2.2.3 Other SOC models and criticality signatures

Besides sandpiles and stochastic branching processes, many other mathematical and physical SOC models have been proposed. In particular, some models are able to keep SOC properties even when some dynamical rules are relaxed. For example, power-laws also appear for a random sandpile model, which topples grains to other random cells instead of neighbor ones, or some particular classes of non-conservative models[4]. Additionally, SOC has been famously employed to study many distinct complex systems such as earthquakes (Olami-Feder-Christensen model (Olami et al., 1992)), solar flares (Lu and Hamilton, 1991), propagation of forest fires (Drossel and Schwabl, 1992), and brain activity (Beggs and Plenz, 2003). Although an extensive discussion of all different model classes and their applications is out of our scope, we briefly mention here a few interesting measurable properties that those systems have in common, which we generally refer to as *criticality signatures*. This concept is particularly important when comparing SOC models to experimental systems where criticality cannot, in principle or due to experimental limitations, be measured.

Many criticality signatures have already been mentioned in this chapter, and, certainly, the most discussed in the literature is the power-law distributions of events' size and duration[5]. Being a direct consequence of a second order phase transition which is present in SOC systems, this scaling has become an indication of critical phenomena. However, it is important to stress that power-law scaling alone does not prove criticality in any system[6]. In fact, power-laws are fairly common in nature and appear for word frequency distribution in specific English novels, number or scientific paper citations, cities' populations, or the diameter of craters in the moon (Newman, 2005). Curiously, those distributions are also alternatively named Zipf's law and/or Pareto distributions in different systems due to historical reasons. Additionally to power-laws, "true" critical systems typically display a $1/f$ scaling of

---

[4] Some models, however, only reach a state near, but not exactly at, criticality. This behavior has been named self-organized quasi-criticality (Bonachela and Munoz, 2009).

[5] Curiously, many natural systems go "beyond" power-laws: huge catastrophic events tend to be more frequent than expected by simple scale-free properties. These particularly rare and large events have been named *dragon kings*, and their origin is still debated in the literature (Sornette, 2009).

[6] We ignore for now the practical consequences of fitting true power-laws and identifying spurious "power-law-like" distributions that are best fit by other functions. See Appendix B for a more extensive discussion on power-law fitting.

power-spectra (Bak et al., 1987; Marković and Gros, 2014), power-law scaling of other critical exponents (for example, the aforementioned ratio between events' duration and size exponents, $\gamma = \frac{\alpha-1}{\tau-1}$), scale-free average avalanche shapes (Friedman et al., 2012; Beggs and Timme, 2012), and must display different behaviors at criticality and at subcritical or supercritical regimes (as we have shown for sandpiles and stochastic branching processes).

The more criticality signatures a natural system displays, the more evidence of a critical point is accumulated, supporting the hypothesis that a theoretical SOC model is indeed a good description. However, rarely can a natural system be shown to be a pure SOC phenomenon, and as more experimental measurements are made, the SOC classification can be weakened or strengthened. Ideally, enough experimental evidence would be gathered to support or undermine each criticality signature, but in practice that is rarely possible. For example, recording in neural circuits cannot be made for every single neuron in the network, which affects virtually every detectable criticality signature (Priesemann et al., 2009), while the tuning of parameters using synaptic antagonists can be rather difficult or imprecise (Beggs and Timme, 2012). It is important, therefore, to understand as precisely as possible the self-organization mechanisms acting on a given system before proposing it as SOC, as well as the limitations of such a description. In the next section, we begin to describe how these mechanisms act in neural circuits and what dynamical rules are known to be present in the brain.

2.3 Self-organization in neural circuits

As one would expect, self-organizational mechanisms in biological neural circuits tend to be more complex and numerous compared to the ones from theoretical toy models. Synapses, for example, are constantly fluctuating not only spontaneously, but also due to plasticity, in response to myriad different stimuli, internal activity, learning processes, or memory formation. Neurons, rather than single cells with a few possible states, interact in multiple different manners depending on their types of ion channels, synapses, functions, neurotransmitters, spike rates, and other factors[7]. The result of these inter-

---

[7] Of course, as the majority of computational studies of neural circuits, we ignore here the roles of other types of cells in the brain, such as glia cells, even though they are far more numerous than neurons in the nervous systems. For a review on different types of glia cells and relatively recent developments on their understanding, see Fields (2009).

actions are macroscopic phenomena we, as humans, observe on a daily basis: muscle movement, decision making, vision and pattern recognition, among others. As they are typically huge systems (the brain has approximately 100 billion neurons (Herculano-Houzel, 2009)) with so many interaction mechanisms and possible resulting behaviors, modeling self-organization in neural circuits is particularly challenging, and there are no obvious order parameters to measure. Instead, at the microscopic level, self-organization dynamical rules have been linked to different types of synaptic plasticity mechanisms, which underlie synaptic efficacy fluctuations and have been widely described in experiments. We review here the key processes underlying synaptic plasticity, which are essential ingredients to understand self-organization towards criticality and the occurrence of criticality signatures in the brain.

2.3.1 Spike-timing-dependent plasticity

The changes in synaptic efficacy due to a neurons' activity are exemplified by the Hebbian theory (or rule, or postulate) (Hebb, 1949), which is often summed up as *cells that fire together, wire together* (Löwel and Singer, 1992). The theory states that if a pre-synaptic neuron A takes part in the firing of post-synaptic neuron B by firing shortly before, the synaptic efficacy from A to B is increased. Thus, repeated firing leads to stronger local wiring, which in turn leads to a higher probability of sequential firing. This proposal was later experimentally verified for spikes occurring inside a limited time window, and has become known as *spike-timing-dependent plasticity* (STDP), which combines the long-term potentiation (LTP) and long-term depression (LTD) of synapses. LTP (LTD) phenomena refer to the findings that very high (low) frequency stimulation of pre-synaptic neurons results in the strengthening (weakening) of the respective synapses and more (less) activity on the post-synaptic neurons (Bliss and Lømo, 1973). This result has since been confirmed in distinct brain areas and animals (Dayan and Abbott, 2001), suggesting a robust match to the Hebbian theory.

The experimental measurement of STDP in different studies (Markram et al., 1997; Bi and Poo, 1998) has shed light on the importance of the spike timing for LTP and LTD effects on excitatory synapses and offered further support to a biological implementation of the Hebbian theory[8]. In fact, the

---

[8] Note that, in principle, LTD/LTD and STDP do not result in the same synaptic efficacy increase/decrease phenomenon. There is, however, evidence suggesting that they

amplitude of excitatory post-synaptic currents (EPSC) has been shown to depend on the precise spiking time, with LTP occurring for positive spike timings (i.e., post-synaptic firing after pre-synaptic firing) and LTD for negative spike timings, as long as they occur inside a time window of tens of milliseconds (Fig. 2.5A). The STDP positive (negative) feedback has the effect of identifying (non) causal relationships, therefore encoding information about activity patterns as a form of unsupervised learning.

Similarly to STDP between excitatory neurons, inhibitory synapses are also affected by spike timings, and the effects of fluctuations in the inhibitory synaptic efficacies (inhibitory pre-synaptic neurons) play an important role in the overall stabilization of network dynamics (Vogels et al., 2013). However, there are some key differences between STDP and inhibitory STDP (iSTDP) (see Haas et al. (2006), Fig. 3). In contrast to STDP in excitatory synapses, iSTDP has very little effect for near zero spike timings, peaking at around 10 ms. Additionally, the strongest changes occur for potentiation after an unsuccessful inhibition (positive spike timings), while weaker depression takes place otherwise, on average. iSTDP has also been proposed to work together with STDP to regulate the overall activity level based uniquely on the spike timing in connected neurons (Haas et al., 2006).

Additionally to the classic, Hebbian STDP rules described so far, other factors are known to influence synapse efficacies. For instance, the amplitude of the fluctuations in the cortex is affected by the presence of neuromodulators, a process which has been named three-factor learning rules (Frémaux and Gerstner, 2016). Neuromodulators typically add another time window in which STDP can modify synaptic efficacies efficiently, linking each synaptic change to inputs or events from the recent past and playing an important role in learning processes. The most studied neuromodulator is dopamine, a neuromodulatory signal produced in the ventral tegmental area (VTA) and transmitted in widely studied pathways that correlate with novelty and received reward (Schultz et al., 1997). In practice, neuromodulation via dopamine in the brain likely aims to bridge the temporal gap between sensory stimulation (which happens at the scale of seconds) and synaptic plasticity (which, as we have mentioned, happens at the millisecond scale). However, whether self-organization by neuromodulation of various STDP mechanisms in recurrent neural networks can achieve improved learning and memory capacities

---

consist in the same general mechanism that aims to improve information encoding while minimizing energy consumption (Yger and Harris, 2013; Krieg and Triesch, 2014).

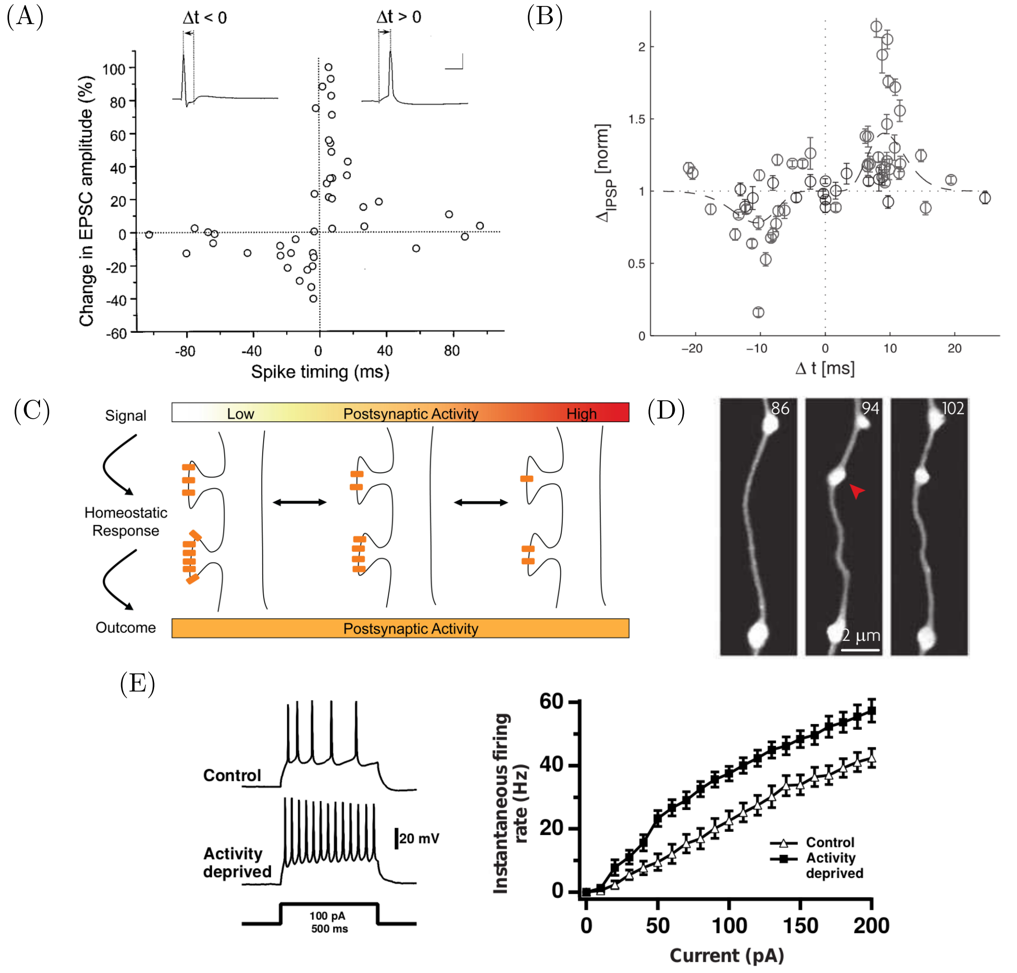


Figure 2.5: **Self-organization mechanisms in neural circuits.** (A) Spike-timing-dependent plasticity (Bi and Poo, 1998): synaptic potentiation ($\Delta t > 0$) and depression ($\Delta t < 0$) based on changes in the excitatory post-synaptic currents (EPSC) as a function of spike timings. Inside a critical time window, synapses with positive timings are potentiated (on average), while synapses with negative timings are depressed. (B) Inhibitory spike-timing-dependent plasticity (Haas et al., 2006): fluctuations of inhibitory postsynaptic potentials (IPSPs) as a function of spike timings for different cells. Synaptic efficacies change very little for $\Delta t \to 0$, and the overall function is asymmetric. (C) Synaptic scaling (Watt and Desai, 2010): Schematic diagram representing changes in the post-synaptic activity, illustrated by the multiplicative scaling in the number of AMPA receptors at synapses (orange rectangles). (D) Structural plasticity (De Paola et al., 2006): time-lapse *in-vivo* imaging (in days — top left corner) of axonal bouton growth (red arrow) for an intracortical mouse axon, suggesting the creation of new synapses. (E) Intrinsic plasticity (Desai et al., 1999): neuronal increased firing as a response to activity deprivation. Left: example of increased firing rate. Right: increase in firing rate for 18 neurons as a function of the input current. All figures were reproduced (or adapted, for (D)) with permission from the respective manuscripts.

remains an open question, and a detailed discussion about its consequences is out of our scope of this thesis.

2.3.2 Homeostasis

Synapses do not fluctuate only due to spike timings, but also as a result of "competition" for limited resources in the system (Turrigiano, 2011), resulting in slightly different self-organization mechanisms. Such competition keeps the activity bounded at a network level and stabilizes a neuron's output under a range of different inputs (Watt and Desai, 2010). The two main homeostatic plasticity mechanisms act, therefore, at different levels: *synaptic scaling* keeps the total input to a neuron fixed, while *intrinsic plasticity* regulates the overall activity level in order to counterbalance variations in the input or recurrent network drive[9].

Synaptic scaling has been observed in cortical, hippocampal, and spinal neurons, both excitatory and inhibitory, which makes it one of the best described homeostatic plasticity mechanisms (Turrigiano et al., 1998; Watt and Desai, 2010). It directly regulates the strength of synapses by altering the number of post-synaptic receptors, thus acting to scale synapses and constrain them to an optimal size range (Watt et al., 2000; Watt and Desai, 2010). This mechanism results in a competition for resources, in which synapses (and therefore also their post-synaptic currents) grow and shrink their relative size (e.g., the number of AMPA receptors) in order to counterbalance the overall input level a neuron receives[10]. Additionally, experimental evidence suggests that this process, differently from STDP, can be multiplicative (Fig. 2.5C) and aid learning and memory processes by modifying only the relative strength differences between inputs (Watt and Desai, 2010). Due to the multiplicative nature of the input scaling, this plasticity mechanism is also referred to as *synaptic normalization* (the name we later use in this thesis), which differentiates it from other types of synaptic scaling.

At the network level, the overall activity is regulated by intrinsic plasticity mechanisms that act on the neurons' excitability at long time scales. This

---

[9] There are also additional homeostatic plasticity mechanisms that act alone or combined in specific brain regions and/or aid in the regulation of the neurons firing rate. For a review, see Watt and Desai (2010).

[10] Experimental evidence supporting a pre-synaptic scaling also exist, due to conservation mechanisms found in the axon, but we do not discuss it here. See, for example, Sabel and Schneider (1988).

activity regulation modifies the neurons' firing rates in response to network changes, thus introducing an activity level control. In particular, neurons have been shown to reduce their excitability when activity deprived (Desai et al. (1999); Fig. 2.5E) and, likewise, increase their excitability when activity increases above a baseline level (Turrigiano, 2011). Interestingly, this process does not have to be local and can potentially affect neurons that are not directly connected or even nearby, as intrinsic plasticity might be chemically regulated by diffusion of specific molecules along the neural circuit (Sweeney et al., 2014). Similarly to synaptic scaling, this overall excitability regulation likely takes place after a competition for resources (diffusive molecules), and one neuron can only become more active when others reduce their activity, in what can be seen as a self-organization process. Importantly, the target firing rates for different neurons do not have to be, in principle, the same, and many different types of neurons are known to fire with different frequencies and/or firing patterns (Izhikevich, 2003). Intrinsic plasticity, therefore, does not dictate which firing rate each neuron should adopt but instead drives such rate towards an overall network level.

2.3.3 Structural plasticity

In addition to spike timing or homeostatic mechanisms, synapses can also be created or removed due to structural changes in neural circuits. This process is known as *structural plasticity* and can be the result of learning, the formation of new memories, or recovery from damage. Structural plasticity is, however, relatively less consistent and robust than other plasticity mechanisms, and its nature is less clear compared to them (Bourne and Harris, 2011). Studies have suggested that this mechanism can be either homeoastatic (Bourne and Harris, 2011) or activity dependent (Holtmaat and Svoboda, 2009), although it is a highly heterogeneous process whose exact function remains an open question. Creation or removal of synapses are typically measured via temporal imaging of axon boutons or dendritic spines' growth or reduction (Fig. 2.5D) and have been observed for neurons in various brain regions (Holtmaat and Svoboda, 2009), suggesting that structural plasticity is potentially ubiquitous in neural circuits, and therefore important for their self-organization.

Finally, we have reviewed concepts and experimental evidence for self-organization in neural circuits via plasticity action. In fact, many of these plasticity mechanisms are active at the same time (although in different time scales) and interact with each other (Abbott and Nelson, 2000; Watt and Desai, 2010). The dynamics of neural systems is shaped by those in a similar (but much more complex) fashion than sandpile models or branching processes, which are regulated by their own self-organization rules. The combined action of spike-timing-dependent plasticity, homeostasis, and structural plasticity, thus, is able to drive and maintain networks at special states, which are potentially the result of selective evolutionary pressures under which plasticity in the brain evolved. But what are those states? Is there a unique dynamical regime towards which biological neural networks are always driven? And, not less importantly, does this hypothetical state display SOC, as theoretical sandpiles or stochastic branching processes do? In an attempt to better understand those questions and their answers, the critical brain hypothesis, which we introduce in the next section, has emerged.

2.4 The critical brain hypothesis

The brain is a system in which self-organization takes place via the interaction between various plasticity mechanisms and activity-dependent processes. In short, the critical brain hypothesis states that the brain, naturally driven by those mechanisms, is poised at a phase transition point, between two types of very different dynamics (Beggs and Plenz, 2003; Chialvo, 2010). This phase transition is, according to the hypothesis, a consequence of a SOC process, including the brain and neural circuits in the list of critical phenomena. The experimental evidence supporting such claim initially came from *in-vitro* experiments: power-laws of events' size and duration have been measured for bursts of activity that spread through the neural circuit and are separated by quiet periods (Beggs and Plenz, 2003). Drawing inspiration from the avalanche nomenclature of SOC systems, this phenomena has been named *neuronal avalanches*[11]. Interestingly, the observed power-law exponents matched the theoretical predictions for stochastic branching processes (-2 for avalanche durations and -1.5 for avalanche sizes), depending on the detection threshold selection and width of time bins (Priesemann et al., 2013). Many experiments have since replicated these findings in other preparations

---

[11] Note that, in general, *neuronal* refers to properties of a single neuron, while *neural* refers to collective properties of a network of many neurons. However, even though avalanches span many neurons, we keep here the historical nomenclature of *neuronal avalanches*.

and brain regions (Mazzoni et al., 2007; Pasquale et al., 2008; Tetzlaff et al., 2010; Lombardi et al., 2012; Friedman et al., 2012; Yang et al., 2012; Shew et al., 2015), and although the initial observation of neuronal avalanches lacked some more detailed statistical treatment of the multielectrode array recordings and power-laws, the hypothesis has since gained momentum and support (see Chapter 4 for a comparison between different experimental setups and examples of observed power-laws).

The measurement of power-law distributed neuronal avalanches raised important questions about the tuning of a complex dynamic system such as the brain. In analogy to branching processes, this critical dynamical regime should keep activity between a subcritical state (when, as we have described, activity reaches zero in a finite number of time steps) and a supercritical regime (in which activity propagates through the whole system indefinitely). Plasticity, as a combination of self-organization mechanisms, is the best candidate for maintaining the brain near this specific point. However, it might seem obvious that a healthy, operating brain constrains its activity to a certain level, avoiding dangerous extreme sub or supercritical regimes (Marković and Gros, 2014). It is surprising, nevertheless, that this is achieved under a range of external stimulation, which can destroy power-laws in theoretical SOC models. In fact, strong external input indeed breaks down power-laws in the visual cortex, but only during a short transient period (Shew et al., 2015), indicating that at least some specific brain regions (such as the early layers of the visual cortex) possess plasticity mechanisms that are capable of maintaining criticality.

Later experiments on systems that receive continuous external input, such as *in-vivo* preparations, only showed criticality signatures for coarse, large-scale measures, while spiking activity, after proper statistical treatment, resembled instead a driven, slightly subcritical regime (Priesemann et al., 2014). Such results suggested a small addition to the original hypothesis, as external input seems to play an important role in determining in which state neural circuits operate. Interestingly, the brain could operate at a slightly subcritical regime when receiving external input in order to avoid supercriticality, which has been linked to epileptic seizures (Meisel et al., 2012). However, such interpretation is not currently a consensus, since other experiments have detected no power-laws in brain activity (Touboul and Destexhe, 2010; Dehghani et al., 2012) and/or suggested that they appear due to other reasons, including a simple occurrence in pure stochastic processes (Touboul and Destexhe (2010); we review these criticisms in more depth in

Chapter 4). The extent to which each experimental procedure, thresholding process, and power-law fitting method affects the observed power-laws is still subject to debate, and thus the origin and function of the recorded power-law scalings remain relatively obscure. For this reason, combined with the fact that power-law scaling alone does not prove criticality, the critical brain hypothesis remains still a hypothesis.

Evidence of SOC in neural circuits is, in some cases, not only limited to power-laws of avalanche events. Other criticality signatures have been detected in particular studies. Examples are the scaling of critical exponents and scale-free average avalanche shapes (Beggs and Timme, 2012; Friedman et al., 2012) and the seemingly necessary level of excitation and inhibition in order to remain at the (supposed) second order phase transition point (Haldeman and Beggs, 2005; Hesse and Gross, 2014). These measurements have not, however, been widely replicated in comparison to the power-law distributions for neuronal avalanches, but are commonly employed as an argument to support the claim that those power-laws are indeed "true" criticality signatures.

Finally, a last claim of the critical brain hypothesis is that, while at criticality, the brain has multiple advantages regarding information processing, and therefore such a state could be the result of evolutionary pressures. Computational modeling of simpler neural networks and branching processes has shown that various information processing capacities are maximized at a critical point (Kinouchi and Copelli, 2006; Shew and Plenz, 2013). Interestingly, those theoretical results do not make it a necessary condition for the whole brain to operate at criticality. For example, hierarchical networks have been shown to display many criticality signatures even when their underlying processes are not critical (Friedman and Landsberg, 2013). Thus, it is conceivable that different brain regions self-organize into distinct dynamical states, some critical and some not, and the general dynamics remain beneficial for information processing. In particular, it would be expected that lower layers that receive direct external input might commonly deviate from criticality, while deeper layers that only receive recurrent network drive are able to maintain a robust critical state (Marković and Gros, 2014).

2.5 Discussion and outline

We have described here how simple theoretical systems with simple self-organization rules might become critical phenomena without external interference, a process also known as SOC. These models reproduce the behavior of different natural self-organizing systems, in which seemingly purposeful patterns arise. We have described in more detail two important examples of critical models: the Bak–Tang–Wiesenfeld sandpile model and stochastic branching processes. Both can be used as analogies to biological neural circuits, and the latter, in fact, is commonly employed in computational models in order to study the critical brain hypothesis. Neural circuits, nonetheless, have rather complex dynamical mechanisms which have only recently been experimentally described, and the interaction among them, as well as their exact function, is subject of ongoing research. Differently from typical SOC models, these plasticity mechanisms can depend on the neuronal properties, activity timing, network structure, or even homeostatic processes, which makes their representation as simple toy models particularly difficult.

The critical brain hypothesis has emerged from the observation of criticality signatures *in-vitro* in the form of power-law distributed sizes and duration of neuronal avalanches, in a direct analogy to branching processes. However controversial due to experimental methods, analyses, and a few contradictory results, this hypothesis has gained popularity to explain brain self-organization and adaptation under different input regimes. As criticality has many benefits for information processing in neural networks, it is reasonable to assume that the brain evolved towards a critical state at the network level. Interestingly, *in-vivo* recordings have shown that such a hypothesis might mean that when receiving external input neural circuits operate at a driven subcritical regime, avoiding dangerous supercritical dynamics. How plasticity mechanisms are able to adapt and maintain different dynamical states under different external inputs is, however, currently unknown.

In the remainder of the thesis, we investigate how criticality signatures might emerge due to plasticity action in recurrent neural networks. First, in Chapter 3, we present a recurrent network model that self-organizes due to biologically inspired plasticity mechanisms. In Chapter 4, we propose a link between criticality signatures and one of the most important brain functions, learning, discussing how both phenomena might arise due to the same self-organization processes. We study how different types of external drive might break down the power-laws in the neuronal avalanche distributions, and reproduce experimental findings on readaptation due to fast plasticity action. Additionally, in Chapter 5, we extend our analysis to another function that has been previously linked to criticality, the memory capacity, and provide examples of applications of self-organizing recurrent neural networks for simple language learning and sentence generation tasks. Last, we summarize our main conclusions in Chapter 6, while suggesting a direction for follow-up studies, both in the field of SOC in neural circuits and of spatio-temporal sequence learning with self-organizing recurrent neural networks.

Chapter 3

Modeling neural activity: self-organizing recurrent neural networks

> *Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.*
>
> George Box

The goal of this chapter is to briefly introduce neural networks, focusing on reservoir computing and sequence models, and to describe in detail the family of models known as Self-Organizing Recurrent Neural networks (SORNs; Lazar et al. (2009); Zheng et al. (2013)). The first half of the chapter is dedicated to the historical developments that motivated the original SORN model and its subsequent variations. We present the implementation of the models' different versions and summarize the most important past results. SORNs are largely used in the remainder of this thesis, and further technical details about their *python* implementation can be found in Appendix A. The source code for the simulations presented in this thesis and its usage instructions can be found online at `https://github.com/delpapa/SORN_V2`.

3.1 A brief historical perspective

The SORN model was introduced by Lazar et al. (2009) in the context of reservoir computing and first employed to study sequence learning tasks. Since then, modified versions of the model have been used in a number of different spatio-temporal learning tasks and as simple models of brain activity, with various degrees of success.

The idea of modeling neural activity using artificial neural networks is, however, by no means new. The first artificial neuron (i.e., a neuron modeled as a mathematical function) dates back to the '40s (McCulloch and Pitts, 1943). Each artificial neuron, or McCulloch & Pitts unit (MCP), combined weighted numerical inputs from different sources and passed them forward to a binary transfer, or activation, function (or, equivalently, compared them to an activation threshold), which then calculated a binary output. This function was inspired by the anatomy of a typical biological neuron, with relatively straightforward analogies (Fig. 3.1): incoming weights multiplying the inputs could be compared to dendrites, the summation of inputs to the soma, and the output transmission to the axon. These simplified MCP neurons are still the building blocks of a wide range of neural network models to this day, including the SORN, and were used in one of the first implementations of an artificial neural network, the Multi-Layer Perceptron (MLP).

3.1.1 Perceptrons

Perceptrons were one of the first attempts to model neural circuits, developed by Frank Rosenblatt in the '50s (Rosenblatt, 1958). The main Perceptron machine, designed for image recognition, employed a combination of MCP units which were implemented directly into custom-built hardware. Such proposal attracted a reasonable amount of attention due to the new parallels with biological neurons, the possibility of adjustment of incoming weights according to different inputs, and the controversial claims regarding its capacities[1]. Using modern notation, the perceptron algorithm can be summarized

---

[1] In 1958, Rosenblatt himself remarkably stated that *"(...) the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence. Later perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech and writing in another language, it was predicted."* (Olazaran, 1996). Although modern machine learning research has partially accomplished some of those goals, perceptrons alone proved insufficient for

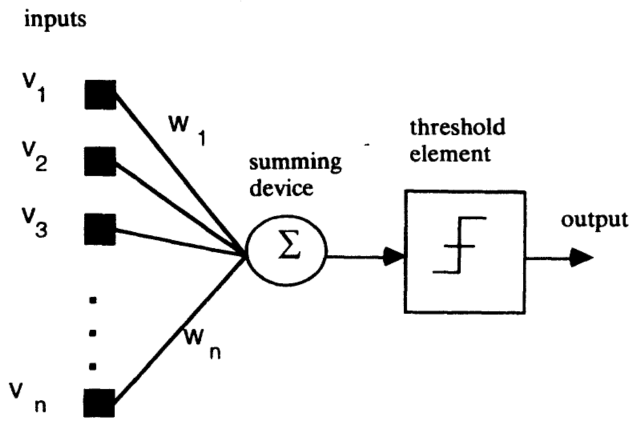


Figure 3.1: **Artificial neurons.** Representation of the McCulloch and Pitts processing unit. Each input v_i is multiplied by an incoming weight w_i, combined in a summing device, and passed forward to a threshold element, which then outputs a single binary value. Reproduced with permission from Olazaran (1996).

as a simple function:

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} w_i x_i + b > 0, \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where x_i are the components of the input vector $\mathbf{x}$, w_i are the perceptron weights, and b is a bias term independent from any input. The weights and bias can be adapted, or learned, for a given problem, and the output of $f(\mathbf{x})$ can be used, for instance, to classify an input instance as a negative or positive example.

Despite seemingly an initially promising algorithm, the single layer perceptron (Fig. 3.2) was still a linear classifier (i.e., it was linear in $\mathbf{x}$), thus being unable to solve problems that are not linearly separable such as the XOR logic function. This shortcoming was demonstrated shortly after their initial proposal, in an influential book by Marvin Minski and Seymour Papert (Minski and Papert, 1969). Given that many interesting learning problems cannot

---

those particularly complex tasks.

be solved purely by simple linear classification, the original enthusiasm for neural networks decayed in the following years, in favor of other approaches towards artificial intelligence (Olazaran, 1996). Interestingly, it was already known that MLPs can overcome this linear classification limitation, but their resurgence only occurred many years later, with the proposal of the backpropagation algorithm (Rumelhart et al., 1986) and the emergence of deep neural network architectures.

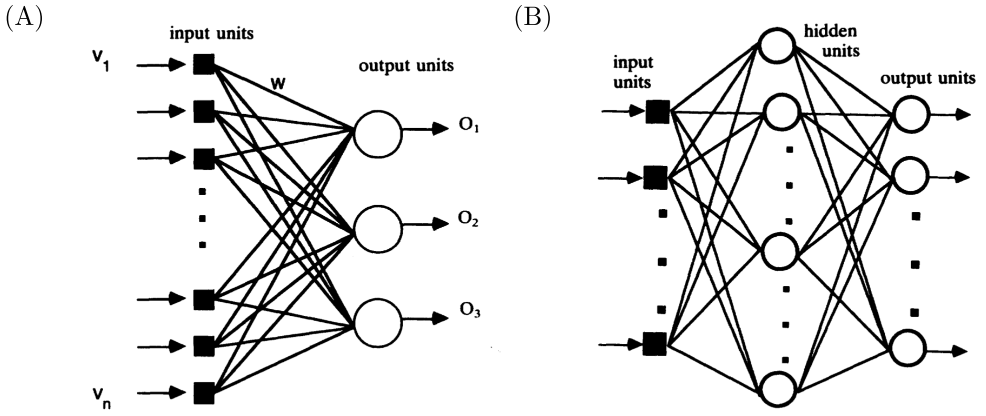


Figure 3.2: **Single and Multi-Layer Perceptrons.** (A) A single layer perceptron. Inputs v_i are multiplied by weights w_{ji}, combined and passed through an activation function. The resulting o_j elements are the perceptron outputs. Reproduced with permission from Olazaran (1996). (B) A Multi-Layer Perceptron (MLP). Combining multiple single layer perceptrons in a row results in a simple feedforward neural network, with input, hidden, and output units. The combination of layers takes place by using the output of a single layer as the input to the next. Reproduced with permission from Olazaran (1996).

In more detail, MLPs consist basically of multiple single layer perceptrons organized in a feedforward manner, where the weighted outputs of a layer are received by the subsequent layer until a final output is calculated (Fig. 3.2B). Each unit i in a given layer may be connected to any other unit j in the next layer by a weight w_{ji}, which is then updated at every training step. These networks are composed of a minimum of three layers: *input*, *hidden*, and *output*. Differently from single layer perceptrons, MLPs combine many nonlinear activation functions, which increase their application from simple binary classification tasks to multi-class classification and regression problems.

Today, MLPs and other feedforward neural networks have become the basis for more complex deep learning models. However, it is relatively straightforward to identify their shortcomings for modeling biological neural circuits. First, networks of neurons are not simply feedforward, showing many examples of recurrent loops (or "back" connections) and intra-layer connections, particularly in the cortex (Douglas and Martin, 2004). Second, the backpropagation algorithm that updates the weights is a non-local learning rule, which requires the computation of precise gradients relative to a predefined loss function. There is no evidence that such processes can occur in the brain at the neural level (Kandel et al., 2000). Third, neurons produce spikes with different firing rates and sometimes precise firing patterns (as in songbirds (Hahnloser et al., 2002), for example), which suggests that their activation functions might have more constraints than a MLP unit. Fourth, synapses are a result of myriad biochemical processes (Kandel et al., 2000), which weights represented by single real numbers are not able to fully capture. Those facts suggest that, in order to model biological neural circuits, additional properties should be at least partially taken into account. Thus, we continue by shifting our focus to tackle the first shortcoming: from the development of pure artificial feedforward networks to more general recurrent models.

3.1.2 Recurrent neural networks

Recurrent neural networks (RNNs) can be seen as a generalization of the feedforward models described previously, in which connections between layers are replaced by connections between any two neurons, as in a directed graph. Those extra connections allow for temporal sequence encoding, as some input information remains stored in their internal state for a number of time steps, a process that gives rise to a memory capacity[2]. Historically, these networks have been used in problems that require temporal represen-

---

[2] Perhaps one of the earliest examples of a recurrent network with a memory capacity is the famous Hopfield network (Hopfield, 1982), which combines binary neurons with Hebbian learning rules and has an interesting associative memory property. However, it suffers from similar training shortcomings as other recurrent neural networks, particularly when compared to biological systems.

tation, such as sequence learning, speech recognition (Sak et al., 2014), and music generation (Boulanger-Lewandowski et al., 2012).

Even though RNNs are able to encode time-dependent information in their internal state, they are classically trained with backpropagation through time, which requires their expansion (or "unrolling") in feedforward-like networks (Werbos, 1988). This process makes standard RNN architectures difficult to train with gradient descent techniques due to gradient vanishing problems, as the size of the "unrolled" network increases exponentially (Bengio et al., 1994; Pascanu et al., 2013). Additionally, large problems become virtually unfeasible due to computational time. To overcome these problems, new approaches have emerged in recent years, based on the introduction of different gating functions. Today, the most widely used architectures are known as Long Short-Term Memory units (LSTMs; Hochreiter and Schmidhuber (1997)) and the Gated Recurrent Units (GRUs; Cho et al. (2014)), both of which were developed aiming for higher performance in relevant problems instead of biological realism[3].

3.1.3 Reservoir Computing

Reservoir computing emerged as a solution to the problem of training complex RNNs models. It is a method in which recurrent connections are randomly generated and kept fixed during training, while only a supervised readout layer is trained for a particular task (Lukoševičius and Jaeger, 2009). The term, coined as an analogy to water reservoirs, broadly refers to two very similar models, which historically appeared in different contexts, namely Echo State Networks (Jaeger and Haas, 2004) and Liquid State Machines (Maass and Markram, 2004)[4]. The analogy relies on the similarity between the be-

---

[3] It is important to also mention one interesting historical example of biologically inspired recurrent neural networks. Balanced networks (Van Vreeswijk and Sompolinsky, 1996), as they have been named, showed that a model combining excitatory and inhibitory neural populations exhibited chaotic behavior, with potential applications for temporal encoding. This model provided an example of development in computational neuroscience, with models that were not constructed aiming only for higher performance in tasks.

[4] In this thesis, the term *reservoir* is used to refer to the recurrent connections and neurons of a network model, excluding the readout layer and input connections. Therefore, following this nomenclature, a reservoir computing model is typically composed of input, reservoir, and readout layer. Additionally, in the following chapters, we use the term *static reservoir* to refer to a reservoir with fixed weights, in contrast to dynamic reservoirs in which weights evolve over time.

havior of wave propagation in a water reservoir's surface and the input in these models: both slowly fade away with time while interacting with other waves or inputs following a nonlinear dynamics. Thus, an important ingredient of these reservoirs is their nonlinear functions, which are known to be extremely beneficial for learning (Huang et al., 2006), even when combined to a simple linear readout layer. As reservoir weights are commonly fixed, training reservoirs turns out to be much faster than more complex RNNs, and the training procedure follows a relatively straightforward recipe (Lukoševičius, 2012). In general terms, past inputs are encoded in the internal state of reservoirs, and due to the recurrent connections "echo" for a number of time steps before fading away. Interestingly, there is some evidence that similar representation decays occur in the brain (Buonomano and Maass, 2009), although cortical circuitry shows non-random connectivity features (Song et al., 2005; Perin et al., 2011). Altogether, these findings raise the question of whether reservoirs are also adequate models for studying biological circuits.

3.1.4 Plasticity induced self-organization

Reservoirs, however powerful, rely on random connections for their spatio-temporal learning abilities. These static random connections limit the number of internal states a network of a given size may store, suggesting that improvements could still be made. For instance, recurrent networks operating near a phase transition state are known to have maximal fading memory capacities (Bertschinger and Natschläger, 2004), a property not observed in random static reservoirs (Del Papa et al., 2019). Additionally, different forms of biologically inspired plasticity mechanisms have been shown to also improve a reservoir's fading memory (Lazar et al., 2007), proposing that insights into brain function might be gained by combining particular plasticity mechanisms with the architecture of reservoirs.

Finally, here is where we reach the SORN model (Lazar et al., 2009). Initially proposed as an improved reservoir for spatio-temporal tasks, the SORN has allowed for a network level description of particular biological neural networks. Those two results alone make the SORN a very interesting model to achieve our goals of understanding the role of self-organization towards criticality in biological neural networks and its consequences on the fading memory capacity. The next section of this chapter describes in mathematical detail how the SORN combines different plasticity mechanisms to update the main reservoirs' weights, a process we henceforward call self-organization.

3.2 The SORN model

In just a few words, SORNs are reservoirs of perceptron-like neurons with dynamic synaptic weights evolving according to biologically inspired plasticity mechanisms (Fig. 3.3). The original model, introduced by Lazar et al. (2009), was developed to study sequence learning tasks and later shown to reproduce a wide range of findings on spontaneous brain activity and the variability of neural responses (Hartmann et al., 2015). As we will discuss later in this thesis (Chapters 4 and 5), such learning abilities result partially from an improved fading memory capacity exhibited by the network, which arises from the combination of plasticity mechanisms. An extended SORN model (Zheng et al., 2013), which incorporates additional plasticity mechanisms and neuronal membrane noise, has been shown to reproduce the distribution and fluctuation patterns of cortical synaptic efficacies, while spontaneously generating synfire chains (Zheng and Triesch, 2014). Additionally, the combination of different plasticity mechanisms can affect the appearance of power-law distributed bursts of activity, commonly associated with healthy dynamics in the brain (Del Papa et al. (2017); Chapter 4). In order to better understand the details and implications of all these findings, we describe in the next sections the models' dynamics and implementation.

3.2.1 Network dynamics

The SORN models consist of a reservoir of N^{E} excitatory and $N^{\mathrm{I}} = 0.2 \times N^{\mathrm{E}}$ inhibitory threshold MCP neurons (McCulloch and Pitts, 1943), whose state at each discrete time step t is described by the binary activity vectors $\mathbf{x}(t)$ and $\mathbf{y}(t)$, corresponding to the activity of excitatory and inhibitory neurons, respectively. Biologically, each discrete time step corresponds to 10 to 20 ms, corresponding to the membrane time constant of biological neurons and the typical scale of spike-timing-dependent plasticity (see below). Both neuron types can be active ("1" state) or silent ("0" state) at each time step depending on their input, membrane noise, and firing threshold. Neurons are connected by synaptic weights w_{ij} (from neuron j to i, by convention) and synapses are allowed between different excitatory neurons (W^{EE}), from excitatory to inhibitory neurons (W^{IE}), and from inhibitory to excitatory neurons (W^{EI}). Connections between inhibitory neurons and self-connections are not included in the models. At each time step t, the network state is updated according to the input each neuron receives and its current threshold. For each neuron i, the following update equations apply:

$$\mathbf{x}_i(t+1) = \Theta\Bigg[\sum_{j=1}^{N^{\mathrm{E}}} w_{ij}^{\mathrm{EE}}(t)x_j(t) - \sum_{k=1}^{N^{\mathrm{I}}} w_{ik}^{\mathrm{EI}}(t)y_k(t) + u_i^{\mathrm{Ext}}(t) + \xi_i^{\mathrm{E}}(t) - T_i^{\mathrm{E}}(t)\Bigg], \quad (3.2)$$

and

$$\mathbf{y}_i(t+1) = \Theta\left[\sum_{j=1}^{N^{\mathrm{E}}} w_{ij}^{\mathrm{IE}}(t)x_j(t+1) + \xi_i^{\mathrm{I}}(t) - T_i^{\mathrm{I}}\right], \quad (3.3)$$

in which $\Theta[.]$ represents the Heaviside step function, which maps the neuronal activations to binary outputs, $u_i^{\mathrm{Ext}}(t)$ is the external input received by neuron i at time step t, and T_i^{E} and T_i^{I} and the excitatory and inhibitory neuronal thresholds, respectively. $\xi^{\mathrm{E}}(t)$ and $\xi^{\mathrm{I}}(t)$ represent the neuronal membrane noise, set to a Gaussian random variable with zero mean and $\sigma^2 = 0.05$ variance unless stated otherwise.

It is important to emphasize here that the membrane noise level ξ can be seen as one of the parameters that regulate the amount of input the neurons receive at each time step. In practice, the noise variance controls the intensity of random inputs a neuron might receive from some other sources not included in the model, thus being essential for many of the results we present in this thesis. For example, as will be shown in Chapter 4, large values of σ^2 result in a network of virtually independent neurons, thus nullifying most of the interesting properties of the SORN models. The inclusion of neuronal membrane noise (introduced in Zheng et al. (2013)), however, has two biological motivations. First, it accounts for the inputs a neuron might receive from other brain areas not included in the model. Second, it accounts for synaptic failure, by providing negative input to some neurons with a small probability and keeping them from firing at a particular time step.

The models are initialized as random reservoirs where each directed connection of the sparse matrix W^{EE} is present with a probability p^{EE}. The remaining weights W^{IE} and W^{EI} are dense matrices in which all the weights

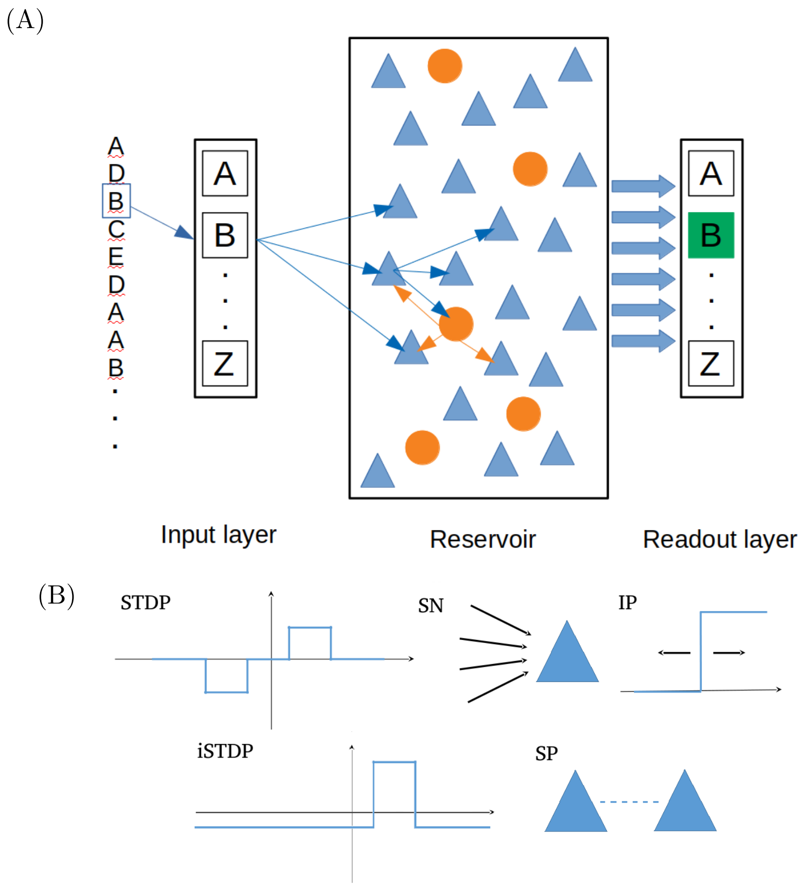


Figure 3.3: **The Self-Organizing Recurrent Neural network model (SORN).** (A) The SORN: a reservoir of excitatory (blue) and inhibitory (orange) neurons receives sequential inputs from an input layer, and a supervised readout layer classifies its internal state for a given task (for example, identifying which input has just been shown to the network). (B) Diagrams of the plasticity mechanisms in the network. The top row shows the three mechanisms in the SORN$_{\mathrm{L}}$: STDP, connections between excitatory neurons are strengthened (y axis) when the post-synaptic neuron is active exactly one time step after the pre-synaptic neuron (positive values in the x axis), and weakened by the same amount when the opposite occurs; SN, the sum of incoming connections to any excitatory neuron is normalized and kept constant over time; IP, the thresholds of the excitatory neurons are updated according to the network activity, increasing when a neuron fires and decreasing when it does not. The bottom row shows the additional two mechanisms incorporated in the SORN$_{\mathrm{Z}}$: iSTDP, the connections from inhibitory to excitatory neurons are strengthened (y axis) when the inhibitions in unsuccessful, i.e., the inhibitory neuron fires one time step before the excitatory neuron (positive x axis) and weakened otherwise; SP, connections between previously unconnected excitatory neurons are added with a small probability while very weak connections are pruned.

are present. All individual weights are drawn from a uniform distribution over the interval $[0, 1]$ and normalized so that the incoming excitatory and inhibitory synapses separately sum up to 1 for all neurons. The excitatory and inhibitory thresholds are initially randomly drawn from the uniform distributions $[0, T^{\mathrm{E}}_{\max}]$ and $[0, T^{\mathrm{I}}_{\max}]$, respectively, and the initial network state $\mathbf{x}(0)$ and $\mathbf{y}(0)$ is randomly selected. Following the implementation from Zheng et al. (2013), the initialization parameters are $p^{\mathrm{EE}} = 10\%$, $T^{\mathrm{E}}_{\max} = 1$, and $T^{\mathrm{I}}_{\max} = 0.5$.

3.2.2 Plasticity mechanisms

The synaptic weights W^{EE} and W^{IE} and the excitatory thresholds T^{E} are subject to plasticity at each time step t, while the weights W^{EI} and inhibitory thresholds T^{I} remain fixed during the simulations. In total, the network employs five different kinds of biologically inspired plasticity, described below (see Fig. 3.3B for simplified diagrams of the plasticity mechanisms action).

Spike-timing-dependent plasticity (STDP)

As a biologically inspired form of Hebbian learning, a discrete model of STDP (Gerstner et al., 1996; Markram et al., 1997; Bi and Poo, 1998) acts on all active excitatory to excitatory synapses, increasing each weight w^{EE}_{ij} by a fixed learning rate η_{STDP} when neuron i fires exactly one time step after neuron j. Conversely, the weight is decreased by the same value if neuron j fires one time step before i. Very small weights ($w^{\mathrm{EE}}_{ij} < 10^{-6}$) are pruned after each STDP update. Formally, both update rules can be combined and STDP is written as a simple update rule, which is applied at each time step:

$$\Delta w^{\mathrm{EE}}_{ij}(t) = \eta_{\mathrm{STDP}} \left[x_i(t) x_j(t-1) - x_j(t) x_i(t-1) \right] \tag{3.4}$$

Inhibitory spike-timing-dependent plasticity (iSTDP)

Similarly to STDP, iSTDP acts on the synaptic weights, but from inhibitory to excitatory ones (W^{EI}). This plasticity rule adjusts those weights in order to balance the amount of excitatory and inhibitory drive that the excitatory neurons receive. This STDP-like phenomenon has been experimentally observed in the cortex (Haas et al., 2006; Vogels et al., 2013) and suggested to be essential for the maintenance of functional cortical circuitry (Vogels et al., 2011). In order to balance the excitatory and inhibitory inputs, each iSTDP step acts as follows. When an inhibition is unsuccessful, i.e., an inhibitory neuron k firing does not prevent an excitatory neuron i firing at the next time step, the weight w_{ik}^{EI}, if present, is increased by $\eta_{\mathrm{iSTDP}}/\mu_{\mathrm{IP}}$, in which $\mu_{\mathrm{IP}} < 1$ represents the mean target firing rate of the network. If the inhibition is successful, i.e., if i is silent one time step after k firing, w_{ik}^{EI} is reduced by a smaller value η_{iSTDP}. In practice, this synaptic weights update regulates the overall network activity, and can be simply written as another update rule:

$$\Delta w_{ij}^{\mathrm{EI}}(t) = -\eta_{\mathrm{iSTDP}} y_j(t-1) \left[1 - x_i(t)(1 + 1/\mu_{\mathrm{IP_i}})\right] . \tag{3.5}$$

Structural plasticity (SP)

In order to compensate for the pruning of excitatory synapses resulting from STDP, SP adds new synapses between previously unconnected excitatory neurons. Specifically, a new synapse is added between a previously unconnected neuron pair at each time step with a small probability p_{SP} and set to a small value η_{SP}. This plasticity rule simulates the constant generation of new synapses observed in both cortex and hippocampus (Johansen-Berg, 2007; Yasumatsu et al., 2008), which occurs even in the adult brain as a result of diverse processes, such as learning new skills or recovering from injuries. In the SORN, nonetheless, the majority of the newly created synapses are quickly eliminated, but a few are strengthened and become part of the network dynamics (Zheng et al., 2013).

Synaptic normalization (SN)

At each time step, after the STDP, iSTDP and SP updates, SN normalizes separately the incoming excitatory and inhibitory synaptic weights of every excitatory neuron, thus regulating the total amount of input it receives while maintaining the relative strengths of the synapses. In the brain, such process avoids uncontrolled growth of any single synapse and has been observed in both excitatory and inhibitory cases (Bourne and Harris, 2011). In particular, this normalization is achieved by the multiplicative scaling of the synapses (Turrigiano et al., 1998; Abbott and Nelson, 2000), which is modeled by applying the following update rule to the weights of W^{EE} and

W^{EI}, separately:

$$w_{ij}(t) \leftarrow w_{ij}(t) / \sum_j w_{ij}(t)\,. \tag{3.6}$$

Intrinsic plasticity (IP)

Last, simultaneously to the updates of the synaptic weights, a variety of homeostatic mechanisms control the neural firing rates, including refractory periods, firing rate adaptation and intrinsic plasticity occurring in different time scales (Desai et al., 1999; Zhang and Linden, 2003; Turrigiano, 2011). The SORN simplify all those processes with a single update rule, maintaining a constant mean target firing rate for each neuron i, $\mu_{\mathrm{IP_i}}$, drawn from a Gaussian probability distribution with mean μ_{IP} and standard deviation σ_{IP}. In practice, the speed of the homeostatic plasticity is regulated by a learning rate η_{IP}, and the IP rule applied to the excitatory neuronal thresholds T_i^{E} can be written as:

$$\Delta T_i^{\mathrm{E}} = \eta_{\mathrm{IP}} \left[x_i(t) - \mu_{\mathrm{IP_i}} \right] \,. \tag{3.7}$$

As can be concluded from the previous paragraphs, the plasticity mechanisms are responsible for adding a number of parameters to the SORN models. Their default values from Zheng et al. (2013) are summarized in Table 3.1. Unless stated otherwise, these were the default values used for all the experiments in this thesis, as long as a particular plasticity mechanism is active.

3.2.3 External input and spontaneous activity

Besides the plasticity mechanisms, another essential feature of SORN models is the addition of external input to excitatory neurons, which allows the model to perform temporal learning tasks (Lazar et al., 2009). Biologically, this input could represent either the direct sensory input received by different cortical regions (for example, the early layers of the visual cortex) or the input received from other brain areas (later layers of the visual cortex). In practice, the external input to neuron i is included in the models via the parameter

Plasticity mech.	**Parameters**	**Default values**
STDP	η_{STDP}	$4 \cdot 10^{-3}$
iSTDP	η_{iSTDP}	$1 \cdot 10^{-3}$
SP	p_{SP}	$1 \cdot 10^{-1}$
	η_{SP}	$1 \cdot 10^{-3}$
SN	-	-
IP	μ_{IP}	$1 \cdot 10^{-1}$
	σ_{IP}	0
	η_{IP}	$1 \cdot 10^{-2}$

Table 3.1: Summary of default values for parameters employed in the plasticity rules.

$u_i^{\text{Ext}}(t)$ (Eq. 3.2). Specifically, external input is described as a temporal sequence of *symbols* during this thesis, where a single symbol is presented to the network at each time step, as in Lazar et al. (2009) and Hartmann et al. (2015). Each symbol provides extra input to a fixed, randomly chosen, and potentially overlapping pool of $N^{\text{U}} < N^{\text{E}}$ excitatory neurons. The length of the input sequence U_{L}, the number of symbols contained in a given input sequence, henceforward called alphabet size U_{A}, the structure of such input sequence and N^{U} are experiment dependent (see Chapters 4 and 5 for details about the experiments and the parameters' numerical values in each case). Whenever possible, we stick to the previously proposed set of parameters from Lazar et al. (2009) and Zheng et al. (2013).

For experiments in which no input is present ($u_i^{\text{Ext}}(t) = 0, \forall i, t$), the SORN's activity is referred to as *spontaneous activity*, in opposition to the *evoked activity* resulting partially from external stimuli. This activity corresponds to the recurrent drive, which originates from a combination of past inputs and spontaneous self-organization due to plasticity action. Importantly, in the brain, spontaneous activity is highly variable and responsible for experimentally measured trial-to-trial variability in a range of tasks (see, for example, Churchland et al. (2010)), a phenomenon also previously measured in SORN models (Hartmann et al., 2015). Such activity, furthermore, possesses space and time structures and strongly differs from pure noise (Ringach, 2009), partially because it is linked to the underlying network connectivity. Therefore, we emphasize here the difference between spontaneous activity in the SORN and noise-driven dynamical regimes.

3.2.4 Training the readout layer

The last building block of the SORN models is the linear readout layer (Fig. 3.3A), which is trained during learning tasks to evaluate the models' performance or make predictions. This layer is trained separately from the main SORN reservoir (which is trained by the plasticity mechanisms) in a supervised fashion, as typically done in reservoir computing (Lukoševičius and Jaeger, 2009). In general terms, the readout layer learns to classify the SORN's activity at each time step when all plasticity is turned off, during a total of T_{train} time steps. The model is subsequently evaluated for another T_{test} time steps, again without the action of plasticity. The model's overall performance is defined as the normalized number of correct classifications (while the error is the normalized number of incorrect classifications). Depending on the task, the classification could be done regarding the current, past, or future input symbols or positions in the input sequence.

There are many well established supervised learning algorithms that perform linear classification efficiently. Importantly, all of them assume that the data (i.e., the internal state $\mathbf{x}$), can be classified by a regression model that is linear in its parameters, which in turn becomes a strong assumption of the SORN models when performing learning tasks. Specifically, we implement the readout by using a logistic regression layer, which has the same size as the number of classification labels (typically the alphabet size or input sequence length). Although this approach is different from the original SORN paper (Lazar et al., 2009), which employed the pseudo-inverse algorithm, we found it faster and computationally more stable for our learning tasks, particularly in the limit of very long input sequences. Both methods, among others, can be used to solve the readout classification problem in reservoir computing, with different assumptions but virtually equivalent results (Lukoševičius, 2012).

3.3 Previous studies and model variations

SORN models have not only been chosen due to their simplicity and easy computational implementation but were also motivated by their previous successes in modeling spatio-temporal learning tasks and cortical dynamics. In this section, we give a brief overview of those past achievements and highlight the differences between the two main SORN variations employed in this thesis: the original model SORN_L (Lazar et al., 2009) and the extended model SORN_Z (Zheng et al., 2013).

Learning tasks

The SORN was first developed to study temporal sequences using a combination of Hebbian learning and homeostatic plasticity (Lazar et al., 2009). The study introduced the SORN_L, which combines three of the five plasticity mechanisms described in the previous section: STDP, SN, and IP, all of which have been shown to improve the fading memory of recurrent neural networks (Lazar et al., 2007). The authors showed how those plasticity mechanisms combined can improve the performance of static reservoirs in different tasks that required the learning of sequential inputs. In particular, when performing a *Counting Task*, which consisted of the repeated presentation of equal length sequences of symbols of the form 'ABB...BBC' and 'DEE...EEF', the SORN_L was capable of outperforming other reservoirs when predicting the next input symbol. Interestingly, the learned internal representations for symbols in the same position were similar, suggesting that the SORN_L indeed learned to differentiate equal symbols in different positions in the sequences, while relying mainly on a local learning rule. The same study also showed that the model quickly reached subcritical dynamics under the input conditions of the same Counting Task, via the measurement of the Hamming distance after small perturbations in its activity. Furthermore, in a follow-up study (Lazar et al., 2011), the same model was shown to learn the statistical structure of its input, suggesting a link between learning via neuronal plasticity and statistical inference.

The learning abilities of the SORN can be better understood by considering the model's response to perturbations and its proximity to criticality. Past studies on the SORN_Z, which adds iSTDP, SP, and membrane noise (ξ) to the original model (see Table 3.2 for an overview of all plasticity mechanisms), have shown that only a fraction of the perturbations become amplified over time, but such fraction decreases as the plasticity driven self-organization takes place (Eser et al., 2014). Interestingly, during spontaneous activity, the authors reported a delay in the amplification of perturbations to the network weights, a *deferred chaos* effect, suggesting that the SORN_Z might not behave like a typical neural network tuned to criticality. Additionally, such a network was shown to spontaneously develop synfire chains (Zheng and Triesch, 2014), allowing for an interplay between faster and slower activity time scales, potentially further aiding its learning abilities. The interaction between plasticity mechanisms and critical dynamics in the SORN is one of the main motivations of our work, and an extended discussion is presented in Chapter 4.

Plasticity mech.	SORN_L	SORN_Z
STDP	✓	✓
iSTDP		✓
SP		✓
SN	✓	✓
IP	✓	✓

Table 3.2: Overview of the SORN model variations and their plasticity mechanisms, as used in this thesis.

Finally, the learning abilities of the SORN_Z have been further tested by an independent research group on a grammar learning task (Duarte et al., 2014). The model was shown to be capable of learning the structure of a Reber grammar (Reber, 1967), performing similarly to humans when judging the validity of a particular grammatical string. This study exemplifies the powerful complex sequence learning abilities of a simple combination of biologically inspired plasticity mechanisms and raises questions about the learning capacity of SORN models in the context of natural language. Such questions are further addressed on Chapter 5, where we investigate the mechanisms underlying its learning abilities and their applications for character-level sentence learning.

Cortical dynamics

Besides the learning abilities, the self-organization due to plasticity mechanisms also has many similarities to cortical self-organization. Zheng et al. (2013) showed that, after self-organization, the overall distribution of excitatory synaptic weights in the $\mathrm{SORN_Z}$ stabilizes as a lognormal-like distribution under particular conditions, matching experimental measurements of excitatory postsynaptic potentials (EPSPs) in the rat visual cortex (Song et al., 2005). Additionally, the fluctuations of those weights qualitatively match the fluctuations of spine sizes in cortical regions (Yasumatsu et al., 2008), while newly created synapses show a power-law distribution of lifetimes, which is also observed in the cortex (Loewenstein et al., 2015). In more biologically detailed SORN models (Hartmann et al., 2016; Miner and Triesch, 2016), a combination of simple plasticity mechanisms have also been able to account for the experimentally observed disproportionate number of bidirectional synapses and for a synaptic efficacy alignment while maintaining many of the properties from the more abstract $\mathrm{SORN_Z}$.

Interestingly, the deterministic plasticity mechanisms of the $\mathrm{SORN_L}$ alone are also responsible for the emergence of important features of neural variability (Hartmann et al., 2015), such as trial-to-trial variability decrease (Faisal et al., 2008) and spontaneous activity alignment with evoked activity patterns (Han et al., 2008). Such results further highlight the importance of abstract SORN models as a link between artificial recurrent neural networks and cortical self-organization modeling. In this thesis, we further explore such a link by looking not only at the learning abilities of the aforementioned plasticity mechanisms but also at their ability to reproduce experimentally observed phenomena.

3.4 Where do SORNs stand today?

Before diving into the results, it is useful to briefly discuss the current state of SORNs in comparison to other state-of-the-art models, both in the fields of computational neuroscience and machine learning. Although a detailed comparison between different models is beyond the scope of our work, other recent studies have extended RNNs and SORNs in other directions, and their differences to the SORNs described here are worth mention.

3.4.1 Beyond binary networks: leaky integrate-and-fire neurons

More biologically realistic SORNs have provided interesting results in the field of computational neuroscience, as recent works have expanded the models from point spiking neurons into the integrate-and-fine domain. The LIF-SORN model (Miner and Triesch, 2016) extended the plasticity mechanisms of the classic SORNs by introducing a dependency on the network's topology and continuous time dynamics. The combination of simple, biologically motivated synaptic, structural, and intrinsic plasticity was able to reproduce, at the same time, important features from cortical circuits, which are not observed in simple static random reservoirs. Those features include the over-representation of bidirectional connections and certain motifs, distance dependent connectivity on local scales, a heavy tailed distribution of synaptic efficiencies, and the power-law distribution of synaptic lifetimes. These results highlight the importance of self-organization in shaping neural circuits, suggesting that many experimental results can be approximated by applying relative simple constraints in a self-organizing neural network, rather than in the typically used randomly wired networks. Additionally, the success of more biologically realistic models based on the same learning mechanisms as the abstract SORNs implies that those mechanisms, which are much easier to investigate in abstract models, are also essential for learning in the brain. This fact further highlights the importance of abstract models with biological constraints for the development of the field of computational neuroscience.

3.4.2 SORNs vs. deep neural networks

In machine learning, deep neural networks have become ubiquitous and are now employed in the most diverse domains of expertise, with fine tuned, task-dependent architectures. Differently from SORNs, deep neural networks are typically trained using backpropagation (Rumelhart et al., 1986) or back-propagation through time (Werbos, 1988). Those are supervised learning rules that rely on the estimation of a general loss function, which is used to update weights via variations of gradient descent algorithms. This training procedure, although computationally expensive, has proved effective for deep feedforward networks and some classes of recurrent networks (the most popular being LSTMs), which today typically consist of hundreds of millions of trainable parameters[5].

In SORNs, as described in previous sessions, learning happens in two stages. First, the main reservoir weights are updated as a consequence of biologically inspired plasticity mechanisms (or *self-organization*), which are activity or topology dependent. Second, a readout layer is trained in a supervised manner to identify the internal state of the reservoir, and the final result is used for classification or generative tasks. Both learning stages have important differences compared to backpropagation. Self-organization is mostly a local learning rule that does not depend on the estimation of general loss functions but on neighboring neurons (STPD) or network activity (IP). Therefore, self-organization is inherently an unsupervised learning rule, which encodes input pattern information into the reservoir's activity and weights. Readout layer training, a supervised learning rule, consists of a single classifier layer, requiring simpler algorithms than backpropagation, such as the pseudo-inverse matrix method or logistic regression. As in the case of more general reservoir computing models, even though computationally faster than backpropagation, these unsupervised and supervised combined methods excel in a different class of problems when compared to deep neural networks[6].

As recurrent neural networks, SORNs are constructed aiming for spatio-temporal learning tasks, which makes a direct performance comparison with LSTMs possible. We discuss this comparison in more detail for a sentence learning and generation task in Chapter 5 but emphasize that our choice for the use of SORNs is motivated by their biological inspiration rather than performance. Although LSTMs have become the common choice for those tasks, they rely on a very different, more complex architecture than self-organized reservoirs. By studying how self-organization affects a reservoir's learning capacity in those tasks, we also provide insights into how reservoirs

---

[5] In practice, due to the huge size, one of the main issues with backpropagation algorithms is the gradient vanishing (or exploding) problem, which results from the propagation of very small (or very large) gradients though many network layers (Bengio et al., 1994). Different architectures have proposed different methods for avoiding or minimizing this problem, but those are not discussed in this thesis as they are not applicable to reservoir computing or SORNs.

[6] Deep learning methods have become more popular than reservoir computing today partially due to the nature of the supervised learning tasks they excel at, such as image classification (Krizhevsky et al., 2012), object detection (Ren et al., 2015), speech recognition (Graves et al., 2013), among others, which take advantage of huge available datasets.

could prove useful for improving current spatio-temporal learning techniques and (potentially) deep recurrent networks architectures, which remains a topic of ongoing research.

3.5 Discussion

We have presented here the SORNs, a family of recurrent neural networks that relies on biologically inspired plasticity mechanisms for self-organization and learning. We described in detail the two main variants of the models, namely the $\mathrm{SORN_L}$ (Lazar et al., 2009) and the $\mathrm{SORN_Z}$ (Zheng et al., 2013), which are employed to model self-organization towards criticality, criticality signatures and spatio-temporal learning tasks in the remaining chapters.

Studying self-organization towards criticality with SORNs has a number of advantages. First, compared to most neural network models, the SORN is a relatively simple one: it has a small number of free parameters to tune, which makes it possible to pinpoint which mechanisms contribute to each observed phenomenon. At the same time, we avoid possible overfitting problems, as complex models with multiple degrees of freedom are more prone to reproduce any desired result with the right set of parameters. Second, SORNs were initially developed as improved reservoirs for spatio-temporal learning tasks, instead of being models previously tuned towards critical dynamics. By identifying criticality signatures in such systems, we highlight the fact that criticality might be an important state towards which self-organizing models with useful learning abilities converge, rather than an artificially chosen state. Third, SORNs stand today between artificial and biological neural networks, with important results in both fields, as described in the previous sections. Critically in neural networks, as a comparison, stands in a similar situation, being a dynamical state with maximal information processing capabilities (Shew and Plenz, 2013) whose signatures are commonly observed in many biological systems (Beggs and Timme, 2012). Last, the plasticity mechanisms and architecture of SORNs allow for insights on brain circuits and application for spatio-temporal learning tasks, a combination that would be impossible with many other models, including more detailed biological networks or deep neural networks.

We often referred to the biological inspiration of synaptic plasticity mechanisms during this chapter, therefore it is important to mention that SORNs can be thought as abstract brain circuits and do not necessarily represent what happens in all layers of different brain regions. Instead of simulating a particular brain region (or the whole brain), we rather focus on general self-organization mechanisms and aim to model criticality and information processing at a network level. As criticality results from the collective behavior of a system and the interaction among its components, such approach is commonly adopted by network models with critical dynamics (e.g., Beggs and Plenz (2003); de Arcangelis et al. (2006); Levina et al. (2007); Poil et al. (2012)), which have various degrees of biological motivation. Certainly, the model's plasticity mechanisms are based on experimental data measured in particular brain regions, which can be used to limit its scope. The ratio between excitatory and inhibitory neurons (5 : 1) is based on data from the sensory cortex (Okun and Lampl, 2008) and hippocampus (Atallah and Scanziani, 2009), as are the studies that motivated the SORN's plasticity mechanisms (Markram et al., 1997; Turrigiano et al., 1998; Bi and Poo, 1998; Haas et al., 2006; Johansen-Berg, 2007; Turrigiano, 2011; Vogels et al., 2013). The addition of neuronal noise, accounting for synaptic failure and inputs from other brain areas, is also compatible with cortical data: the SORN reservoir can be seen as a one layer recurrent neural network, in comparison, for example, to single layers from the visual cortex. Last, the experimental data we aim to reproduce, power-law distributed bursts of activity, have been widely observed in cortical cultures (Beggs and Plenz, 2003; Friedman et al., 2012; Priesemann et al., 2014; Shew et al., 2015), even in blood-oxygen-level-dependent (BOLD) measurements (Tagliazucchi et al., 2012; Shriki et al., 2013), and recently in the whole brain dynamics (Ponce-Alvarez et al., 2018). Therefore, even though SORN's draw inspiration mainly from dynamics in the cortex and hippocampus, they may provide more general insights.

In summary, SORNs possess properties that are essential for investigating the interaction between experimentally observed criticality signatures and learning abilities. Being relatively simple models, they allow for a systematic analysis of their plasticity mechanisms while maintaining useful learning abilities. Naturally, these models might become more general and/or be improved in the future, either by new experimental findings or more powerful learning algorithms, given that they are simple, but not simpler than necessary for our study of abstract circuits.

Chapter 4

Criticality meets learning: neuronal avalanches in spontaneous and evoked activity

> If you try and take a cat apart
> to see how it works, the first
> thing you have on your hands is
> a non-working cat.
>
> Douglas N. Adams

The brain does a surprisingly good job controlling its overall activity level. Such a task is particularly difficult as it requires precise tuning of a huge number of parameters in order to keep the whole system in a healthy and behavioral useful dynamics (Beggs and Timme, 2012). This achievement is even more impressive due to the wide range of different input intensities the brain is capable of processing and learning from without losing its stability. To cope with such variety, specific adaptation mechanisms have evolved to maintain the activity bounded in a healthy regime, avoiding epileptic seizures and long quiescent periods. But how do these mechanisms work? Are there any other functional roles of this particular, *critical*, activity state? In this chapter, we use self-organizing recurrent neural networks (SORNs) to propose a novel link between criticality signatures (in the form neuronal avalanches) and learning, an essential brain function. We provide an extended analysis of different mechanisms and conditions in which this link holds, by studying both spontaneous and evoked activity cases. Most results presented here have been previously published and can be found in our recent paper (Del Papa et al., 2017), while some discussions also appear in Del Papa et al. (2019). Parts of the text of this chapter have also been taken from the same publications.

4.1 Neuronal avalanches and criticality in neural circuits

In theoretical neuroscience, a popular hypothesis claims that biological neural circuits operate near a special state, which is capable of maintaining long term stable dynamics (Beggs and Plenz, 2003). This hypothesis, discussed in more detail in Chapter 2, is today commonly known as the Critical Brain Hypothesis (Beggs and Timme, 2012), as the dynamical state is poised at a phase transition point, between a *subcritical* regime, in which the activity decays and most system units end up in a quiescent state, and a *supercritical* regime, in which activity is amplified over time and most units become constantly active. In order to always operate near such a state, brain circuits would require continuous and precisely tuned adaptation (or self-organization) mechanisms, as many different input conditions and intensities might quickly cause shifts in the current dynamical regime and destabilize the whole network. Thus, the study of criticality in the brain must not only address the question of *whether* the brain is critical, but also *how* it remains critical in an ever-changing environment (Haimovici et al., 2013; Plenz, 2013).

Investigating criticality in the brain, as expected for any biological complex system, is far from an easy task. In typical physical dynamical systems, a critical point can be measured by considering an *order parameter*, which measures how a given system property varies as a function of a *control parameter*, tuned by some outer mechanism. In general, small changes in a control parameter result in small changes in an order parameter; criticality occurs at phase transition points[1], where sudden fluctuations appear and different, unique macroscopic properties might emerge. Such a state also

---

[1] Note, furthermore, that phase transitions can be of first or second order (see section 2.1.1).

commonly marks a shift from ordered to disordered dynamics, as in the classical Ising model example. Neural circuits, however, have no self-evident order or control parameters, as myriad interactions happen continuously between different types of neurons while input is received from varied sources. The relationship between macroscopic measurable properties and microscopic mechanisms is still not fully modeled or even understood[2]. Macroscopic measurable properties linked to a critical brain dynamics, in analogy to the magnetization in the Ising model, are currently considered open questions, although healthy, stable brain dynamics (Meisel et al., 2012) and maximal information properties (Shew et al., 2011; Shew and Plenz, 2013) are commonly proposed as observable quantities. Interestingly, and maybe expectedly, it is nearly impossible to systematically tune a control parameter in living neural circuits, emphasizing the need for other forms of criticality detection and analysis.

Given all the challenges described above, how can criticality be actually measured in the brain? As seen in Chapter 2, based on experiments with toy models and theoretical arguments (Chialvo, 2010; Beggs and Timme, 2012; Hesse and Gross, 2014), critical systems are expected to show chaotic behavior and scale free properties for some particular quantities such as dynamic correlation coefficients and local domain sizes. In particular, for some classes of self-organizing dynamical systems, those scale free properties appear in the size and duration of perturbation events, as observed in the Abelian Sandpile Model, which describes theoretical piles[3] of sand (Bak et al., 1987). As an analogy, similar signatures of criticality can be sought in other complex systems and point to, but not prove, dynamics near a critical point. Typical criticality signatures relate to chaotic behavior, including response to perturbations or Lyapunov exponents approximately equal to zero, and to scale free properties, including power-law distributions of size and duration of activity or perturbation events. In the case of neural circuits, the second of these signatures, power-laws of activity distributions, can be more easily (or at least

---

[2] In fact, even relatively "simple" quantities such as the number of neurons in the brain for different species have only recently been measured (Herculano-Houzel, 2009). Although very complex brain models are currently under development (Markram, 2006), their applicability is subject to intense debate, since very complicated models with a huge number of free parameters can, in principle, fit any desired system.

[3] It is interesting to highlight again that such theoretical models only hold for relatively small experimental piles of sand, breaking down in the limit of large real ones (Held et al., 1990).

less problematically) inferred from activity measurements. In fact, power-law distributed bursts of neural activity have become the most common result suggesting criticality in the brain, on which the critical brain hypothesis relies. Those bursts of activity have been named *neuronal avalanches*, in an analogy to critical sandpile models and sand avalanches, and have gained popularity in the last decade as criticality signatures, after the first convincing evidence that they can be found in neural populations (Beggs and Plenz (2003); Fig. 4.1).

Similarly to the general criticality ideas, the proposal of systems that tune themselves, or self-organize, to a critical state also started in physics and was later borrowed for the description of neural circuits after neuronal avalanches were observed (Hesse and Gross, 2014). The formalization of the main theoretical framework was proposed many years earlier than the measurement of neuronal avalanches, and was known as *self-organized criticality* (SOC; Bak et al. (1988)). SOC in networks has been first demonstrated to occur by simple rewiring rules (Bornholdt and Rohlf, 2000), which raised the question of whether a similar essential process could occur in complex neural networks as a result of more sophisticated and realistic self-organization rules. Years later, measurements of neural activity suggested that such hypothesis had at least some experimental backing: multielectrode array measurements of local field potentials (LFPs) showed that bursts of activity in organotypic cultures of slices of rat cortex have power-law distributed sizes (the number of active electrodes, Fig. 4.1A) and potentials (the sum of LFPs, Fig. 4.1B; Beggs and Plenz (2003)). Moreover, such power-laws showed an exponent of -1.5 for a range of time bin sizes (Fig. 4.1C), which was compatible with theoretical predictions of critical toy systems (Harris, 2002). This finding was the first of many studies that confirmed the existence of neuronal avalanches in many types of neural activity measurements, in different brain regions, including both *in-vitro* and *in-vivo* experimental setups. These experimental results are the basis for our investigation of neuronal avalanches as criticality signatures in SORNs. Therefore, we describe the most important studies in more detail in the next sections, before presenting our main results.

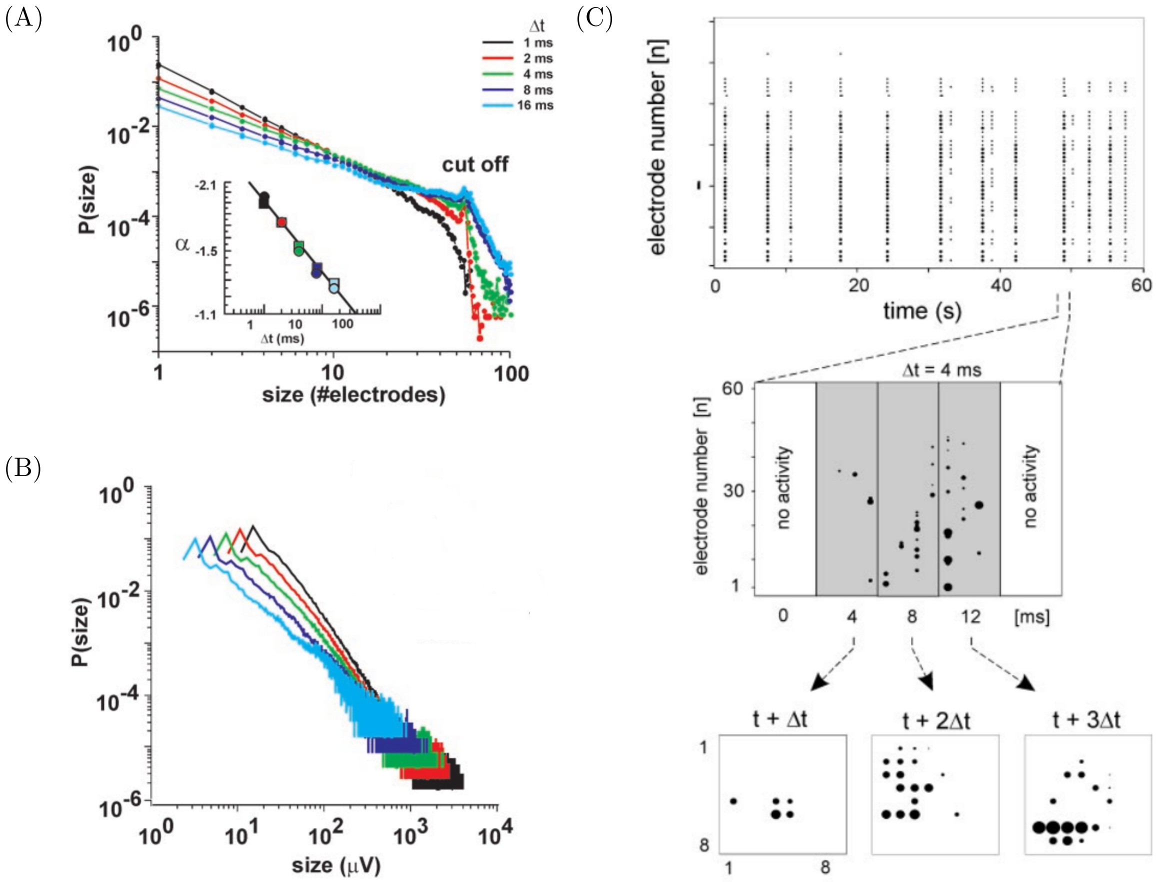


Figure 4.1: **Neuronal avalanches in cortical circuits, from Beggs and Plenz (2003).** (A) Power-law distributions for avalanche sizes, based on the number of activated electrodes, for different time bins. α shows the exponent for each time bin, and the cut-off shows the maximum number of electrodes that recorded activity in the experiment ($n_e = 60$). (B) Probability distribution of avalanche sizes based on summed LFPs from each electrode, for different time bins. (C) Time binning procedure for avalanche recording. Raster plot (top) shows bursts of activity, which could be divided in time bins of different length (middle), representing different electrodes in a square multielectrode array (bottom). Figures have been reproduced ((A) and (C)) or adapted ((B)) with permission from Beggs and Plenz (2003).

4.1.1 Power-laws and criticality signatures in experimental data

After the initial finding of neuronal avalanches in cultures of rat cortical slices (Beggs and Plenz, 2003), many studies sought further experimental evidence to support or undermine the critical brain hypothesis. Following the initial LFP measurements in cortical networks, power-laws have been observed in very different *in-vitro* circuits, such as cultures of cortex (Friedman et al., 2012; Lombardi et al., 2012; Yang et al., 2012), ganglia and hippocampus (Mazzoni et al., 2007), dissociated rat cortical neurons (Pasquale et al., 2008), and developing cortical networks (Gireesh and Plenz, 2008; Tetzlaff et al., 2010), supporting the hypothesis that neuronal avalanches are ubiquitous in neural circuits, independently of their architecture and/or function. Furthermore, many of these measurements reproduced other characteristic properties of SOC systems (see Fig. 4.2 for a selection of these results), such as power-law exponents close to the theoretical expected value of -1.5 (up to a cut-off resulting from the finite size of the systems) for event size distributions (Harris, 2002), a fixed ratio between the exponents of power-laws for event size and duration (Sethna et al., 2001), and the scaling of avalanche shapes (Kuntz and Sethna, 2000). Based on these results, the critical brain hypothesis has gained momentum and support, although important criticisms should be taken into consideration when discussing the validity of the aforementioned experimental results (see next sections and the discussion at the end of this chapter).

Furthermore, different biological mechanisms have been found to cause *in-vitro* networks to deviate from a state in which criticality signatures appear. For instance, alteration of GABA_A neuronal receptors in order to decrease inhibition leads to a supercritical network state (Beggs and Plenz, 2003; Mazzoni et al., 2007; Gireesh and Plenz, 2008; Pasquale et al., 2008; Yang et al., 2012), while an increase in the relative number of inhibitory neurons yields subcritical dynamics (Chen et al., 2010). Excitation, however, acts in an opposite manner: a decrease in excitation by modification of AMPA (Shew et al., 2009; Yang et al., 2012) or NMDA (Mazzoni et al., 2007; Gireesh and Plenz, 2008) receptors results in avalanche distributions that are compatible with subcritical dynamics. Thus, *in-vitro* systems showed that the balance between excitation and inhibition is essential for the maintenance of power-law distributed neuronal avalanches, a potential candidate for a control parameter, and might aid the emergence of functional brain

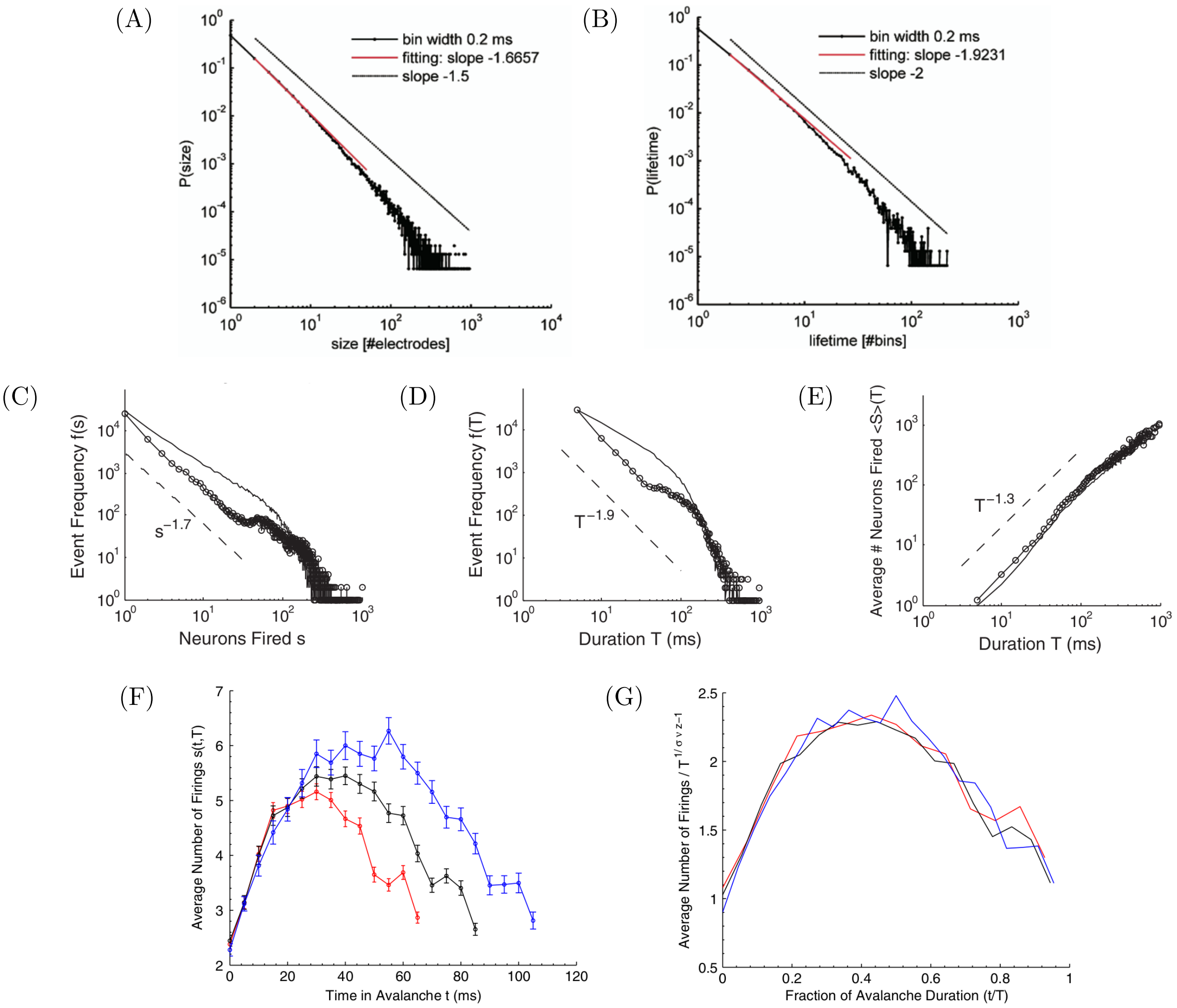


Figure 4.2: **Experimental evidence of criticality signatures in various circuits.** (A), (B) (Pasquale et al., 2008): neuronal avalanches for cultures of dissociated rat cortical neurons and respective power-law fits; (A) size distribution (with slope close to −1.5); (B) duration distribution. (C), (D), (E), (F), (G) (Friedman et al., 2012): neuronal avalanches for a single sample of cortical culture with individual neuron level recordings; (C) size distribution (power-law with exponent τ); (D) duration distribution (exponent α); (E) theoretical prediction for a system near criticality and experimental results for the exponents ratio $(\alpha - 1)/(\tau - 1)$. (F) Example of avalanche shapes in terms of average number of firings (size) as a function of duration; (G) the collapsed avalanche shapes. Figures reproduced with permission from the respective manuscripts. (C)-(G): *Reprinted figure with permission from Friedman, N. et al. (2012): Universal critical dynamics in high resolution neuronal avalanche data. Physical review letters, 108(20):208102. Copyright 2012 by the American Physical Society.*

networks (Hellyer et al., 2016).

In-vivo results, however, tell a slightly different and more complex story (see Fig. 4.3 for a graphical summary of this story). On the one hand, evidence of neuronal avalanches has been found in coarse measures of cortical activity in different animals, such as negative LFP in awake monkeys (Petermann et al., 2009), voltage imaging in mice after (but not during) anesthesia (Scott et al., 2014), and brain blood oxygenated level dependent (BOLD) signals, functional MRI, and magnetoencephalography (MEG) in humans (Kitzbichler et al., 2009; Poil et al., 2012; Tagliazucchi et al., 2012; Shriki et al., 2013). On the other hand, spiking activity in awake animals has failed to show criticality signatures. First, in the cat cortex, not only power-laws are absent, but the $1/f$ frequency scaling of the power-spectra, another criticality signature (see Chapter 2), is not consistent with the expected curve for critical states (Bedard et al., 2006). Second, avalanche analysis in cats, monkeys, and humans showed no power-laws (Dehghani et al., 2012; Priesemann et al., 2013) and suggested a slightly subcritical regime instead (Priesemann et al., 2014). Such disparity implied that healthy neural networks might be able to self-organize towards different dynamical states with potentially different functions, depending on the overall network state. Thus, a better understating of the mechanisms underlying this state dependent self-organization could, for a start, shed light on these seemingly contradictory experimental results, and potentially explain the observed discrepancies.

Interestingly, a given animal behavioral state also seems to have an effect on its brain dynamical state (Hahn et al., 2017). Although power-laws seem consistent across the sleep-wake cycle and during anesthesia (except for spiking avalanches (Dehghani et al., 2012)), important differences have been observed in rats (Ribeiro et al., 2010), cats, monkeys (Priesemann et al., 2014) and humans (Priesemann et al., 2013). In anesthetized rats, for example, other signatures of criticality, such as the $1/f$ spectra scaling, collapse, while they can be observed in the same animals when freely-behaving (Ribeiro et al., 2010). Furthermore, the average size of avalanches has been found to differ among slow wave sleep, wakefulness, and rapid eye movement (from large to small, respectively; Priesemann et al. (2013)). Such differences could be the result of different levels of external stimulation to the system, which varies for each behavioral state, or even an indication that brain circuits self-organize towards different dynamical states for different conditions. The latter hypothesis has some experimental backing: deviations from criticality

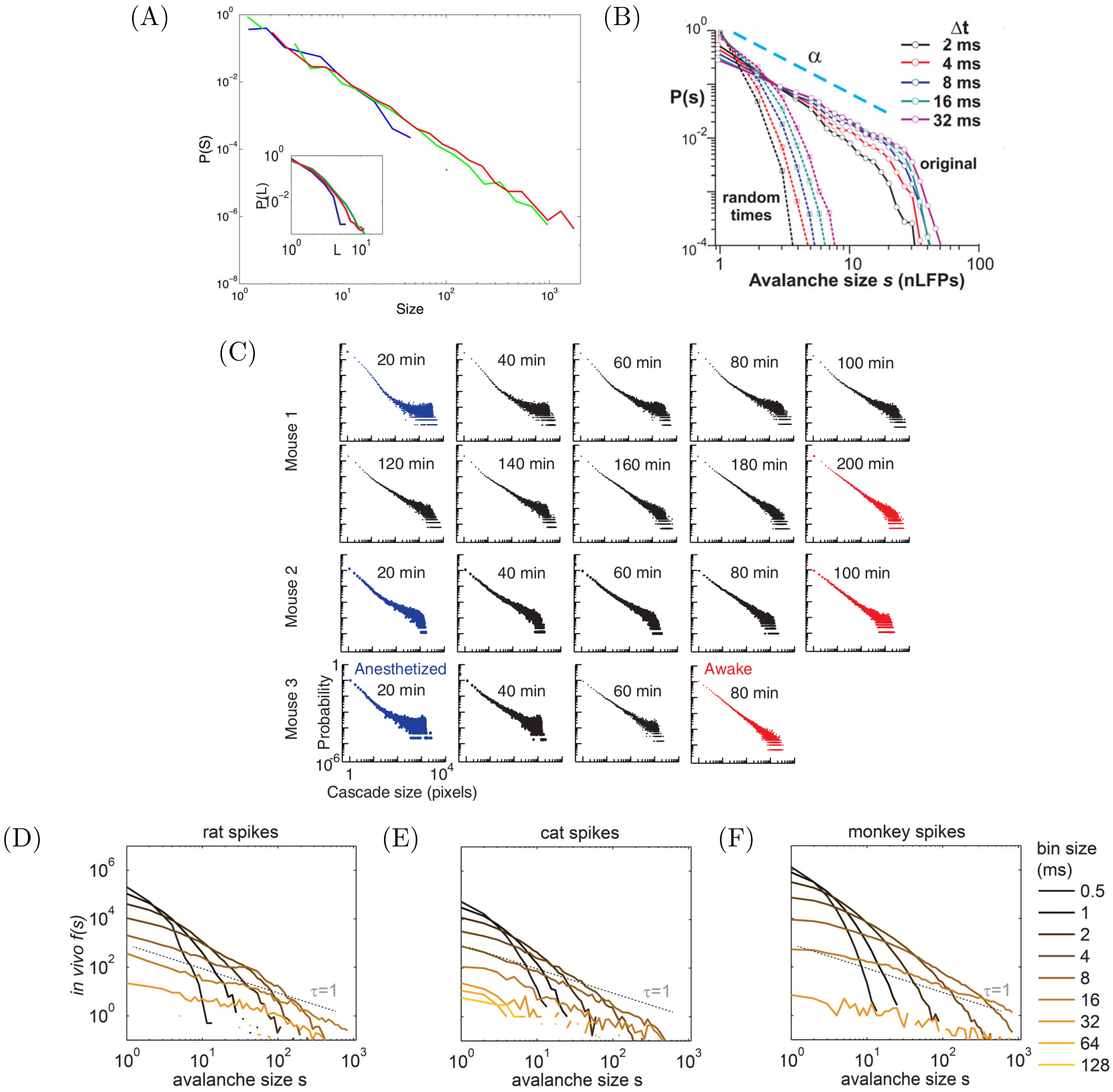


Figure 4.3: **Avalanches *in-vivo*.** (A) (Tagliazucchi et al., 2012): avalanche size (S) and duration (L) distributions for different thresholds of detection (colors) of BOLD activity in humans. (B) (Petermann et al., 2009): avalanche size distributions (negative LFPs) in awake monkeys for different time bins (not observed with shuffled times). (C) (Scott et al., 2014): examples of avalanche size distributions in voltage imaging of mouse cortex recovering from anesthesia. Power-laws break down for anesthetized animals (blue), but are recovered when they become fully awake (red). (D), (E), (F) (Priesemann et al., 2014): avalanche size distributions in the hippocampus of awake rat (D), visual cortex of anesthetized cat (E), and prefrontal cortex of awake monkey (F). They all lack power-laws independently of the time bin size, resembling instead a subcritical model. All figures have been reproduced or adapted (C) with permission from the respective manuscripts.

are related to certain pathologies, as criticality signatures break down during epileptic seizures when the system resembles supercriticality instead (Meisel et al., 2012). Therefore, brain circuits would benefit best from a near critical, but slightly subcritical, dynamics, as they would then avoid small deviations to epileptic regimes while keeping some of criticality's functional gains (Priesemann et al., 2014; Massobrio et al., 2015).

The hypothesis that different levels of external stimulation are at least partially responsible for the different observed dynamical states, should not, however, be ignored. While they receive a larger range of external inputs[4], awake, freely-behaving animals lack the typical separation of time scales expected for SOC systems (Bak et al. (1988); see also Chapter 2), and activity measurements might only be able to detect combined and entangled avalanches, which lack the expected "pure" power-law distributions (Priesemann et al., 2014). Additionally, since external input is continuously received, the brain might never precisely reach a critical point (Bonachela et al., 2010) but operate nearby instead[5], in an extended critical-like region corresponding to a Griffiths phase (Moretti and Muñoz, 2013). Note that this argument does not invalidate the hypothesis that neural circuits operate at a slightly subcritical regime to avoid epilepsy, but can be seen simply as a complementary explanation for all experimental results mentioned here so far. The evolution of brain adaptation mechanisms might have resulted in a network system that, under a wide range of external inputs, self-organizes to operate close to criticality while avoiding dangerous supercritical regimes.

4.1.2 *Ex-vivo* experiments: the role of external input

Neural circuits in awake, behaving animals typically receive a large number of different external inputs (i.e., sensory inputs). Given the importance of those inputs to the adaptation of critical and near critical dynamical systems, experiments have compared how neural networks behave before, during, and after the onset of external stimulation, by measuring *ex-vivo* activity (Shew et al., 2015; Clawson et al., 2017), and suggested that sensory adaptation

---

[4] Incidentally, simple artificial networks tuned to criticality have optimal dynamical range, i.e., they can detect and process information from a large range of input intensities (Kinouchi and Copelli, 2006).

[5] The near critical state in non-conserving systems has been called apparent criticality or *self-organized quasi-criticality* (Bonachela and Munoz, 2009), in contrast to true criticality. In this thesis, we opt for the near criticality nomenclature, for simplicity.

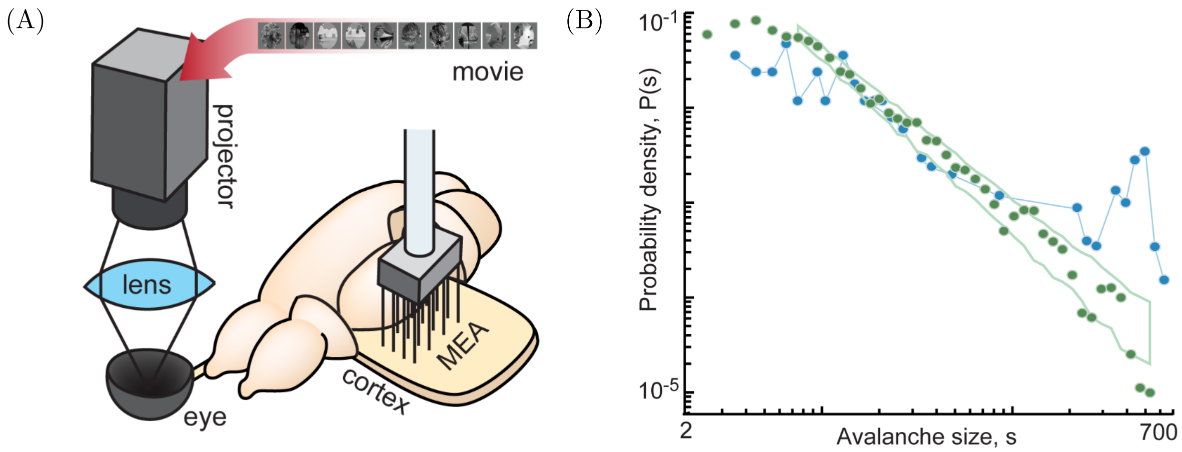


Figure 4.4: **Adaptation tunes cortical dynamics back towards a critical dynamics after short transient (Clawson et al., 2017).** (A) Experimental setup for *ex-vivo* recording in the turtle visual cortex (whole brain brain plus eyes). Movies were projected onto the retina while a microelectrode array (MEA) recorded LFPs. (B) Avalanche size distributions for the transient up to 1 second after input onset (blue) and after the readaptation period (green — green lines show the 5% - 95% probability percentiles for samples drawn from a pure power-law). The transient distribution exhibits a peak for large avalanches, which is a typical property of supercritical systems. Figures reproduced with permission from Clawson et al. (2017).

itself is a self-organizing mechanism responsible for maintaining criticality in the cortex. In particular, LFP and spike recordings were made in the turtle *ex-vivo* brain (whole brain plus eyes), while visual input in the form of a movie was projected into the retina of the preparations (Fig. 4.4A). The results supported the hypothesis of adaptation towards a critical regime: although criticality signatures are lost shortly after input onset, this transient is quickly overcome, and the power-law distributed neuronal avalanches return in the next few hundreds of milliseconds (Fig. 4.4B).

The transient period, although short, revealed an interesting system property. The size of avalanches almost immediately increased (see the peak at the blue distribution in Fig. 4.4B), as more activity was present in the cortical network, a feature typically observed in supercritical systems. Thus, it was possible to conclude from those studies that internal cortical mechanisms acted to bring activity down quickly, possibly to keep it constrained to a near critical, healthy regime, as discussed in the previous section. Furthermore, an increase in the stimulus discrimination was detected only for regimes in which the power-laws appeared, even though stimulus detection was slightly reduced (Clawson et al., 2017). This result, combined with experimental indications that sensory dynamic range is maximized near criticality in the rat whisker system (Gautam et al., 2015), offered further evidence that important brain functions are improved, or even optimized, at criticality.

4.1.3 Modeling neuronal avalanches and criticality

Measuring the improvement of functions in brain circuits for a given dynamical regime is not an easy task. Although the network dynamics can be controlled, for instance, via the onset of external input (Gautam et al., 2015; Shew et al., 2015) or by balancing the amount of excitation and inhibition in the network (Haldeman and Beggs, 2005; Beggs and Timme, 2012; Hesse and Gross, 2014), those approaches can be applied *in-vitro*, while most interesting high-level functions can only be measured (if at all) in *in-vivo*, in awake animals. Additionally, as discussed previously, investigating criticality typically requires the precise tuning and measurement of parameters that are either extremely costly, difficult, or even impossible in living animals. For these and other reasons[6], modeling neuronal avalanches has become an important tool to investigate what would the brain, indeed, gain by operating exactly at a critical point, and how this can be achieved with some biological constraints.

In neural network models, critical dynamics has been widely proposed to arise from ongoing synaptic plasticity action. Power-law distributions of avalanche sizes have been observed in networks of different complexity and plasticity mechanics, ranging from stochastic spiking neurons (Brochini et al., 2016) and simple activity-dependent dynamic synapses (de Arcangelis et al., 2006; Levina et al., 2007, 2009; de Andrade Costa et al., 2015) to spike-timing-dependent plasticity (STDP) (Meisel and Gross, 2009; Uhlig et al., 2013) and a combination of short and long-term plasticity mecha-

---

[6] Of course, modeling a physical system also results in important insights and many useful ideas about how the real system behaves, given that the model is kept simple but good enough. Naturally, this approach is widely accepted in theoretical physics, but less obvious in the study of complex biological systems, where myriad different entities continuously interact. A discussion about how simple models help to improve science in general is, however interesting, beyond the scope of this thesis.

nisms (Stepp et al., 2015). Interestingly, SOC and other similar dynamical states (for example, a "quasicritical" dynamics near a non-equilibrium, Widom line (Williams-García et al., 2014)) can be achieved by the same networks with different tuning conditions and parameters. As mentioned in the previous paragraphs, a popular control parameter is the overall excitatory and inhibitory balance (Shew et al., 2011; Lombardi et al., 2012), a tuning mechanism that has been widely experimentally observed (Beggs and Timme, 2012; Hesse and Gross, 2014). In particular, the level of inhibition seems to be determinant for achieving criticality in randomly connected networks (Neto et al., 2017). Most of these systems and networks have, therefore, been designed and/or tuned to show criticality signatures or critical points, but a unified theory linking self-organization mechanisms to their biologically relevant functions, such as learning and memory, is still to be developed (Haimovici et al., 2013; Plenz, 2013). All these models, however, raise an important argument that should not be left unmentioned: the critical brain hypothesis is, even if not verified, at least plausible.

As one would expect, simpler models are more prone to reproduce criticality signatures not only because they are easier to tune, but because critical dynamics is more evident in systems with simpler dynamics due to the separation of time scales between input, internal activity, and adaptation mechanisms, as in the case of the original sandpile model (Bak et al., 1987) or branching processes (Beggs and Plenz, 2003). While the exact physical and biological mechanisms which the brain employs to arrive at and maintain criticality during learning and development remain unknown, insights have been gained by looking at a variety of models. After those toy models, most numerical studies simulate a network of identical units or neurons, whose activity is regulated by different adaptation rules. Popular choices are activity dependent and/or relative spike timing rules (Hesse and Gross, 2014), in the form of activity-dependent rewiring (Bornholdt and Rohlf, 2000; Tetzlaff et al., 2010), Hebbian learning (de Arcangelis et al., 2006) (but see also anti-Hebbian local learning rules (Magnasco et al., 2009)), short-term plasticity (Millman et al., 2010; Levina et al., 2007, 2009), and, on the more biologically realistic side, STDP (Shin and Kim, 2006; Meisel and Gross, 2009; Rubinov et al., 2011). Given the relatively large number of simplified mechanisms resulting in criticality, the latter seems to be a fundamental property of self-organization mechanics rather than a simple consequence of particular implementation details, suggesting it is a robust process.

The same robustness argument can, in addition, be applied to the type of neuron these models are built with. From systems composed of point, two state neurons inspired in simple branching processes (Beggs and Plenz, 2003) to leaky integrate-and-fire (LIF; de Arcangelis et al. (2006); Meisel and Gross (2009); Rubinov et al. (2011)) and Izhikevich neurons (Teixeira and Shanahan, 2015), and even one example of Hodgkin-Huxley neurons (Shin and Kim, 2006), criticality signatures and scale-free dynamics have been shown to appear. Differently from the case of self-organization mechanisms, however, the computational power can often be a limiting factor for the choice of neuron model: distributions of neuronal avalanches, for example, require a huge number of simulation time steps to be observed, and very complex neuron models are typically impractical. As criticality in neural networks is a dynamical state, a true critical point should be relatively independent of the details of the system's units, which explains why authors typically opt to study its properties and functions with more abstract and computationally tractable neuron models.

As previously mentioned, although criticality can appear in very simple networks, it is important not only to understand *how* it appears in the biological brain, but also *why* would be advantageous for the brain to maintain or not such state. Even though the balance between excitation and inhibition seems a very plausible control parameter to explain the first question, it not only has been shown not to be a necessary condition in artificial critical networks (Jost and Kolwankar, 2009) but also has a poorly understood functional role (Meisel et al., 2012; Lombardi et al., 2012; Hesse and Gross, 2014). Computational models can shed light into this problem via yet another approach: by investigating which biologically useful functions are improved in critical systems. Given that self-organization mechanisms in the brain evolved to optimize information processing, the hypothesis of criticality in the brain can only hold true if a critical state is, indeed, better than other possibilities when processing information. Thus, before addressing the major criticisms of the critical brain hypothesis and finally present our results with SORN models, we discuss which information processing abilities have been shown to be optimized (or maximized) in critical networks.

4.1.4 Maximal information processing

Intuitively, an information processing system performs best when avoiding two extreme scenarios. First, evidently, a circuit should not have a "null" dynamics, i.e., dynamical regimes in which there is no activity at all should be avoided as they do not change over time, and thus do not convey any information. Second, a system composed of always active, or randomly activated, units has similar shortcomings, as no input information can be encoded. Furthermore, experimental recordings have repeatedly shown that the amount of activity in the brain at any single time point (or time bin) is rather limited and only a small number of neurons are simultaneously active[7]. Thus, there must be at least one particular middle level of activity that improves, or ideally maximizes, information encoding compared to extreme, impractical scenarios.

Ultimately, what are the benefits for neural circuits to operate near criticality? So far, we have been mostly using a generic reference to "information processing" or "encoding" when referring to its (supposedly) useful properties, but it is important to be more precise when describing them. In fact, many studies have shown that particular quantities improve in critical neural networks when compared to the same networks tuned to other dynamical states, and we briefly discuss some of them in this section (for a complete review, but lacking the most recent results, see Shew and Plenz (2013)).

Dynamical range Dynamical range[8] is the range of distinguishable stimulus intensities by the activity, or population response, of a system. This property is extremely important, for example, for cortical systems, which receive and initially process external inputs. On the one hand, a subcritical system is capable of responding to weak stimuli due to activity propagation among neurons. Since this activity propagation increases as the system approaches supercriticality, stronger stimuli can also be encoded. On the other hand, a supercritical system gets easily saturated and should not be able to distinguish between large stimuli. This property was verified in externally driven branching processes (Kinouchi and Copelli, 2006), by tuning them to and away from criticality via their branching parameter and measuring

---

[7] Typical neuronal firing rates depend on the particular animal, particular brain region, and are not fixed even for a single neuron under a particular repetitive task (Hartmann et al., 2015). In the *in-vivo* cortex, for example, neurons are estimated to fire with a frequency that varies from less than one Hertz to tens of Hertz (Roxin et al., 2011). As one spike duration is in the scale of milliseconds (but shows many variations depending on neuron type, synapse type, among other factors; Zhang and Linden (2003); Turrigiano (2011)), one may immediately see that neurons are not always active.

[8] Although *dynamical range* is the original nomenclature (Kinouchi and Copelli, 2006), which we follow here, this quantity is also called *dynamic range* in the literature.

their response ratio between strong and weak inputs (Fig. 4.5A). Interestingly, this is one of the rare functionalities that have also been suggested in experiments: by chemically controlling the balance between excitation and inhibition in cortical cultures (and assuming that intact slices are in a critical state because they show power-laws), a similar measure of response ratio was maximized at criticality (Shew et al., 2009). Although their electric stimulation was systematically varied, this particular experiment measured criticality as a deviation from a pure power-law distribution of avalanche sizes, which made their conclusions dependent of the avalanche definition of criticality[9]. Nonetheless, there seems to be a link between the excitation and inhibition balance and optimal dynamical range in cortical networks.

Mutual information between input and response. Mutual information (Dayan and Abbott, 2001) quantifies the ability of a network to encode and transmit information about stimuli to its neural response. The more a system is capable of transmitting useful information about inputs, the better it is at perceiving, and possibly understanding, complex environments. This property is maximized at criticality both in cellular automata (Li et al., 1990) and between different neurons in neural networks (Greenfield and Lecar, 2001). Furthermore, learning and plasticity rules aimed at maximizing mutual information resulted in networks that have been shown to display both neuronal avalanches (Tanaka et al., 2009) and dynamics near criticality (Shriki, 2003). Last, on the experimental side, the encoding of amplitude patterns of electrical stimuli in cortical *in-vitro* cultures is maximal for critical states (Shew et al. (2011); Fig. 4.5B). As in the case of the dynamical range, cortical slices were tuned by the chemical control of the excitatory and inhibitory synapses, and the distance to criticality was measured based on the deviation from power-law distributed neuronal avalanches. Therefore, the same considerations apply, and this can only be considered an experimental observation of maximal mutual information in case the avalanches, indeed, are proof that these circuits were in a critical state. A particularly important application of mutual information in neural networks is their capacity of representing visual inputs (Shriki and Yellin, 2016). After being trained with visual inputs, networks should converge to a representational state, which

---

[9] For a discussion about the possibility of different critical states, one of which not necessarily distinguishable by power-law distributed neuronal avalanches, see the last section and discussion of this chapter.

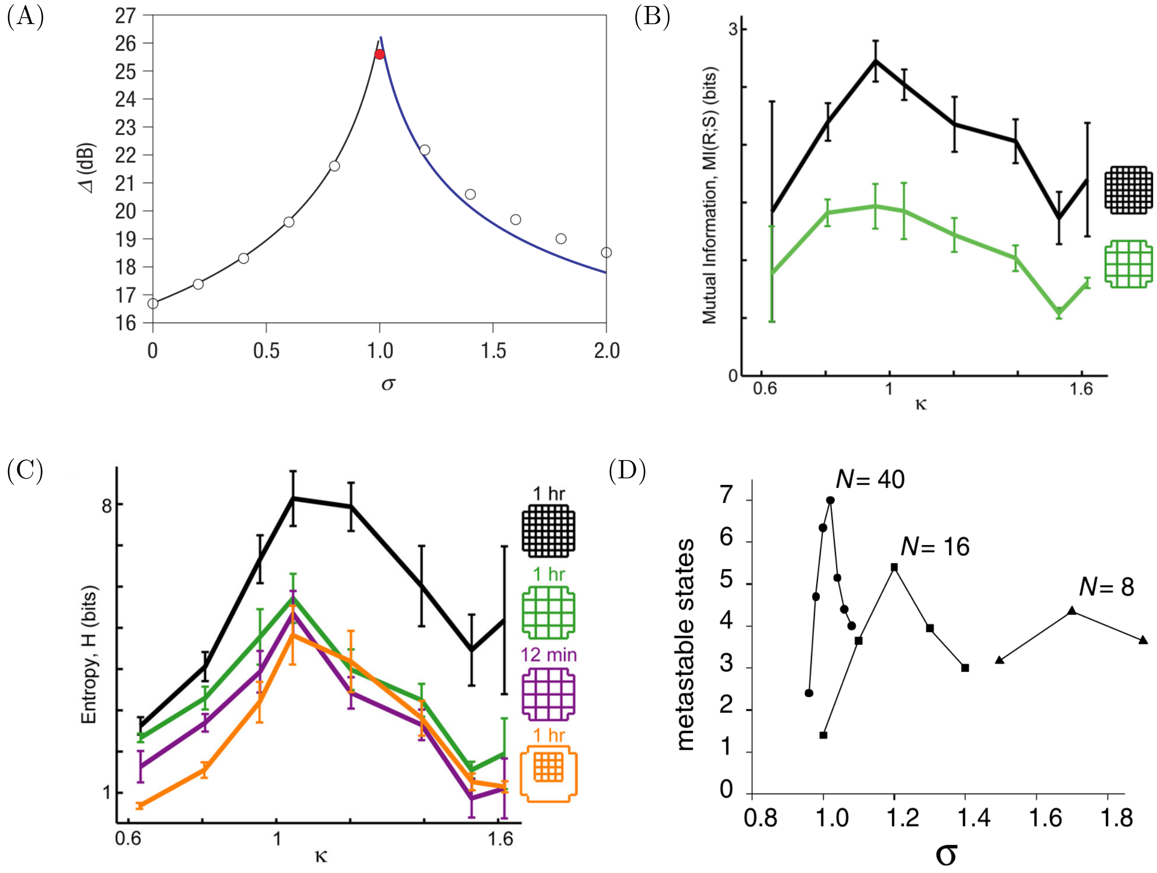


Figure 4.5: **Benefits of critical systems for information processing.** (A) (Kinouchi and Copelli, 2006): dynamical range (Δ, a measure of the response ratio between strong and weak inputs) is optimal for a critical branching parameter (σ) in externally driven branching processes. See the original paper for details on the estimation of Δ. (B), (C) (Shew et al., 2011): (B) mutual information, the mutual entropy between patterns of electrical stimuli and recorded response, and information capacity (entropy) are maximal for networks exhibiting power-law distributed neuronal avalanches ($\kappa \approx 1$, where κ is a measure of the deviation from power-laws) for different recording electrode array geometries and recording times. See the original paper for details on the estimation of the mutual entropy. (D) (Haldeman and Beggs, 2005): information capacity for branching processes (branching parameter σ) is maximal during criticality ($\sigma = 1$), for big enough systems (N nodes). The number of metastable states is a measurement of the repeating activity patterns in the system (again, see the original paper for numerical and analytical details). All figures were reproduced with permission from the respective manuscripts. (D): *Reprinted figure with permission from Haldeman, C. and Beggs, J. M: (2005). Critical branching captures activity in living neural networks and maximizes the number of metastable states. Physical review letters, 94(5):058101. Copyright 2005 by the American Physical Society.*

can be more or less accurate, and from which the input can be, in the ideal case, reconstructed accurately. Remarkably, this convergence becomes much slower during phase transition points, in a process known as *critical slowing down* (Scheffer et al., 2009). By training a network in order to maximize the mutual information between input and output, a recent study (Shriki and Yellin, 2016) showed that both visual input orientation representation (in terms of a population vector (Georgopoulos et al., 1989)) and convergence time undergo a phase transition when the system is tuned, via adjustment of weights, to criticality. Although this phase transition has been shown only in a model trained with gradient descent (thus not biologically realistic), it suggested that those properties might generalize to other networks also tuned towards criticality, and, in case the critical brain hypothesis holds, the whole brain.

Information capacity. Information capacity measures the number of patterns that can be stored in a network's activity. In other words, it is the entropy (Shannon, 1948) of the system. Many neural networks have been shown to have maximal information capacity at criticality, by measurements of the number of activity patterns in branching processes (Haldeman and Beggs (2005); Fig. 4.5D), networks of boolean neurons (i.e., directed graphs; Rämö et al. (2007)), and even cortical slices (Stewart and Plenz, 2008). Interestingly, by defining activity patterns based on electrode recordings, the information capacity is also maximized when power-law distributed neuronal avalanches appear (Stewart and Plenz (2008); Shew et al. (2011); Fig. 4.5C). Last, cortical sensory circuits seem to optimize their performance by maximizing their response entropy with respect to a given input probability distribution (Dan et al., 1996; Rieke and Warland, 1999). This adds one more experimental argument to support that a state which maximizes information capacity is a desired and possibly the best "choice" of dynamical state for neural circuits.

Memory capacity. A last, but essential, information processing property of neural networks is their fading memory (sometimes referred to as working memory): their ability to retain information about recent inputs in their activity (Maass et al., 2002). Such property is a determinant for a network's total memory capacity, which might have a direct link to its learning abilities. Interestingly, the fading memory capacity scales approximately logarithmically with the network size for systems tuned to criticality, but slower for other systems (Bertschinger and Natschläger, 2004). Notably, however, such finding has only been observed for randomly connected, static networks, for which no study has reported neuronal avalanches or other experimentally observed criticality signatures. A potential co-occurrence of neuronal avalanches and the logarithmic scaling of the fading memory capacity in the same system might shed light on the connection between different "properties" (or possibly types) of criticality. We further discuss those properties, their consequences, and explore the applications of systems with improved fading memory capacities in Chapter 5.

Given that criticality seems to improve so many important brain functions, any deviation from a critical dynamics might result in potential brain dysfunctions. As described before, epileptic regimes carry signatures of supercriticality (Meisel et al., 2012), which arguably damage most, if not all, the information processing abilities described here. Additionally, neural circuits operating far from a critical regime might result in abnormal reactions to sensory stimulation and limited neural activity patterns when compared to healthy, near critical systems. One important hallmark of these dysfunctional systems can be speculated to be the loss of balance between excitation and inhibition, as it is a common way of chemically controlling the network regime of *in-vitro* circuits. In fact, this argument has been used to link deviation from critical dynamics to autism (Shew and Plenz, 2013), although no conclusive evidence has been found *in-vivo*. In summary, all these results point out that operating near a critical point would be very beneficial for the brain from an information processing perspective, and evolution might have found mechanisms to assure that in most cases, neural circuits take advantage of those functional benefits.

4.1.5 Criticality criticisms

So far we have presented multiple arguments and experimental evidence to support the critical brain hypothesis, or at least indicating that the brain operates close to criticality. It is, however, important to mention that this topic is still subject to debate in the neuroscience community and no consensus has been reached, mostly due to the difficulty of measuring reliable and general criticality signatures or proving their plausibility in different neural circuits. Beyond simply taking sides in this discussion, we hope to shed light on the topic by exploring SORN models with biologically inspired plasticity mechanisms and proposing a novel link between criticality signatures and learning abilities. Therefore, we mention here the most common criticisms against the critical brain hypothesis and address later their applicability to SORN models and our main results.

The main concern when inferring criticality from neuronal avalanches is that power-law distributions are not unique to critical systems and can be explained by other factors (Marković and Gros, 2014; Touboul and Destexhe, 2010). In fact, even other criticality signatures such as the $1/f$ frequency scaling do not necessarily reflect a SOC or critical state (Bedard et al., 2006; Bédard and Destexhe, 2009). Many other mechanisms can result in power-law-like distributions of activity events, including stochastic processes (e.g., noise convolved with Poison processes or a thresholded Ornstein-Uhlenbeck model (Touboul and Destexhe, 2010)), filtered neural activity (Bedard et al., 2006), or even multiplicative noise (Sornette, 1998). In fact, power-law distributions of events are fairly common in nature (Reed and Hughes, 2002) and arise generically for a class of large systems that lack any fine tuning (Schwab et al. (2014); see also Chapter 2). Thus, any criticality claim cannot be only supported by power-laws, and other indications of critical points must be observed if one wants to show that a neural network is at a critical point. Indicators include, for example, the mathematical relationship between size and duration exponents (Beggs and Timme, 2012), the observation of a control parameter that tunes the system's dynamical regime (Haldeman and Beggs, 2005) or the collapse of avalanche shape as a function of relative duration (which suggests a scale-free system; Friedman et al. (2012)). Although those indicators can be observed in computational models, most experimental setups are not capable of measuring more than a few of them together with the power-laws, which justify the latter's designation as signatures, but not proofs, of criticality.

Besides not being a clear proof of criticality, the measurement of power-law distributed neuronal avalanches may suffer from yet another drawback: the power-law fitting process (Clauset et al., 2009). Power-laws are, by their mathematical nature, difficult to tell apart from heavy-tailed or exponential distributions and require rather sophisticated statistical tests, which are not always employed on experimental datasets (Newman, 2005; Clauset et al., 2009; Marković and Gros, 2014). In fact, many studies have argued that some experimentally observed power-laws do not hold after careful analysis (Bédard and Destexhe (2009); Touboul and Destexhe (2010); but see also criticism of their conclusions regarding neuroscience datasets: Beggs and Timme (2012)). To make matters more complicated, experimentally observed power-laws are truncated, as the neural systems have finite size, and do not typically spread among many orders of magnitude, which make them more prone to misclassification. Given the high possibility of false positives, in this thesis, we follow the recommended approach to fit power-laws (Clauset et al., 2009): a combination of maximum likelihood estimators for the power-law exponents and a comparison of the goodness of fit between different possibilities of distributions. For a more detailed description of this method, see Appendix B.

In experimental data, additional issues might affect the inference of criticality from power-laws. First, neuronal avalanche distributions might depend heavily on the event detection threshold or even time bin sizes (Pasquale et al., 2008; Touboul and Destexhe, 2010). Importantly, however, most results on neuronal avalanches do not depend on the bin size and have carefully compared multiple bin sizes for neuronal avalanche events (e.g., Beggs and Plenz (2003); Tetzlaff et al. (2010) — see also Fig. 4.1), showing that although such criticism demands an additional statistical analysis, it does not invalidate the experimental evidence suggesting criticality in the brain. Second, as experiments are typically done via multielectrode array recording or other less spatially precise methods, a subsampling effect must be taken into consideration when treating the data (Priesemann et al., 2009, 2014). This effect is the result of the array geometry (or equivalent for other recording methods): since the activity is not recorded for every single neuron, a single avalanche can display artificial pauses in case the array geometry is not taken into account. Even pure SOC theoretical models are misclassified in case of subsampling, when activity is not measured in every single cell or unit (Priesemann et al., 2009), showing that corrections must be made to take into account the subsampling geometry. In the subsampled case, for instance, the distribution of events' size and duration might not necessarily follow a power-law, and any event measurement depends both on the subsampling geometry and the intrinsic model dynamics, which complicates the interpretation. Recent methods and estimators have been proposed in order to extract useful information from this subsampling effect (Nonnenmacher et al., 2017), including the system's distance to criticality itself (Wilting and Priesemann, 2016), paving the way to future, more precise experimental insights on the critical brain hypothesis. It is important to highlight that subsampling issues are essentially an experimental feature and, although computational models can be also subsampled, all their neuronal states can in principle be accessed at any point in simulation time, which dismisses the need for any additional corrections.

Last, a smaller but important point about neuronal avalanches *in-vivo* and in externally driven computational models should be raised. Due to the nature of those systems, input and internal dynamics co-occur, resulting in the lack of separation of time scales ((Priesemann et al., 2014); in contrast to traditional SOC models (Bak et al., 1988)). As a consequence, different avalanches are not separated by pauses and occur at the same time, encompassing the same neurons and overlapping, and can be mistakenly detected as one event. A possible solution to disentangle avalanches in those cases is the introduction of detection thresholds, which artificially separate overlapping avalanche events. Although there is currently no formal theory about the effects of thresholds on events' size and duration distributions, computational studies have shown that power-laws still occur in thresholded critical systems (Poil et al., 2012), but the distributions depend on the threshold choice to some extent, as they do in experiments[10] (Hesse and Gross, 2014). This is the case for SORN models: as their internal dynamics never stops due to plasticity action, event detection thresholds must be introduced as additional parameters to our neuronal avalanche measuring method. We discuss in the next sections our threshold choices, their robustness, and how these choices might affect the neuronal avalanche distributions and our conclusions. We highlight, nonetheless, that this process has become common practice for systems with no separation of time scales, both in computational models and experiments (Priesemann et al., 2014; Hesse and Gross, 2014).

4.2 Spontaneous activity: self-organization towards criticality

In this section, we describe the existence of neuronal avalanches in the spontaneous activity of models of the SORN family and discuss which mechanisms might underlie their occurrence after self-organization. In particular, we focus on the plasticity mechanisms and neuronal membrane noise level, which

---

[10] This problem is amplified when combined with subsampling effects, a case which we do not discuss here. See Priesemann et al. (2014) for a explanation of this and similar cases.

can be alternatively interpreted as input from other non-modeled circuits. We argue that even though SORNs might not always operate at criticality, their self-organization mechanisms might bring them towards a regime in which power-law distributed avalanches appear in their spontaneous activity and that those mechanisms are important to understand how the same processes might occur in the brain. In fact, this is one of the reasons why SORN models were chosen for our computational experiments with criticality signatures: by combining biologically inspired plasticity and self-organization mechanisms with relatively simple neurons, SORNs show a reasonable level of abstraction to approach the criticality signatures detection problem, while still allowing for simulations to be run in a computationally feasible time.

4.2.1 SORNs revisited

Self-organizing recurrent neural networks (SORNs; Lazar et al. (2009)) consist of a reservoir of excitatory and inhibitory neurons, whose state at each discrete time step is described by two binary vectors, $\mathbf{x}(t) \in \{0,1\}^{N^{\mathrm{E}}}$ and $\mathbf{y}(t) \in \{0,1\}^{N^{\mathrm{I}}}$. Those vectors correspond to the activity of excitatory and inhibitory neurons, respectively. Connections between excitatory neurons (W^{EE}) and from inhibitory to excitatory neurons (W^{EI}) are shaped by binarized forms of spike-timing-dependent plasticity (STDP), or inhibitory spike-timing-dependent plasticity (iSTDP), by homeostatic mechanisms, namely synaptic normalization (SN), and by structural plasticity (SP), depending on the variation of the model. Additionally, SORN models also include a form of intrinsic plasticity (IP), which regulates the neuronal firing thresholds and may combine their internal recurrent neuronal drive with external input and/or membrane noise. Details of all these mechanisms, their functions, and their mathematical formulations are described in Chapter 3. In this chapter, we refer mostly to one SORN variation: the $\mathrm{SORN}_{\mathrm{Z}}$ (Zheng et al., 2013), which has additional plasticity mechanisms compared to the original model, the $\mathrm{SORN}_{\mathrm{L}}$ (Lazar et al. (2009); see Table 3.2).

For the case of spontaneous activity, we reimplemented the $\mathrm{SORN}_{\mathrm{Z}}$ model due to its richer spontaneous dynamics, large number of biologically inspired plasticity mechanisms (five in total — STDP, iSTDP, SN, SP, and IP), and neuronal membrane noise. Additionally, this model has been shown to reproduce, under certain conditions, features of cortical dynamics, such as the lognormal-like distribution of excitatory weights (Zheng et al., 2013), making it an adequate cortical model in which to investigate self-organization and criticality signatures. Our simulations kept the same set of default parameters as the original model (Zheng et al., 2013) whenever possible.

In our reimplementation of the SORN$_Z$ model (see Appendix A), its spontaneous activity showed three different self-organization phases regarding the number of active excitatory to excitatory synapses when driven only by Gaussian membrane noise (Fig. 4.6A). After being randomly initialized, the number of active connections decreased quickly as a result of the STDP pruning action (the *decay* phase) and reached a minimum at around 10^5 time steps, before slowly increasing due to SP action (*growth* phase) until stabilization after around two million time steps (*stable* phase), when only minor fluctuations were present. Importantly, this result was equivalent to the self-organization phases observed in the original model implementation (Zheng et al., 2013), which provided some validation to ours.

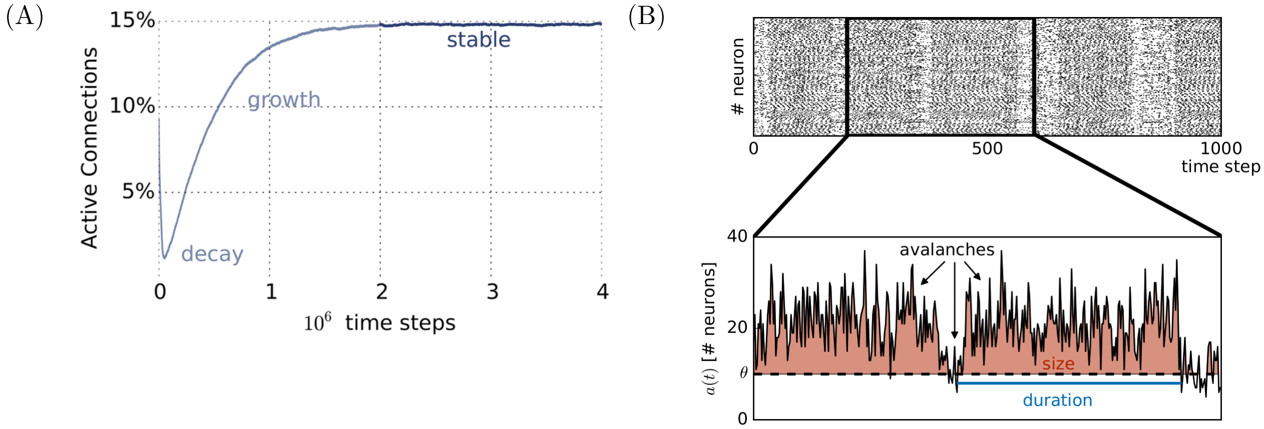


Figure 4.6: **Self-organization phases and thresholded neuronal avalanches.** (A) Fraction of active connections as a function of simulation time in the SORN$_Z$, starting from a random connected graph with 10% of active excitatory to excitatory connections. The model exhibits three self-organization phases: *decay*, *growth*, and *stable*. Neuronal avalanches were observed in the latter. Figure reproduced with permission from Del Papa et al. (2017). (B) Raster plot of the SORN$_Z$ spontaneous activity $a(t)$ in the stable phase, after self-organization, and neuronal avalanche event definition via activity threshold θ. A single avalanche starts when the activity goes above θ and lasts as long as it stays above this threshold. Shaded red areas indicate the size of avalanches and the blue line indicates the duration of a single avalanche event.

In order to avoid possible transient effects, we concentrated our analyses only on the stable phase, discarding the first 2×10^6 time steps. In this sense, we measured neuronal avalanche distributions in the regime into which the SORN self-organizes driven only by membrane noise and its own plasticity mechanisms, diminishing any possible influence of the random initialization of the synaptic weights. Interestingly, during this post-transient phase, bursts of asynchronous activity could be observed, suggesting indeed potential neuronal avalanches (Fig. 4.6B).

As pointed out in the previous sections, SORNs are fundamentally different from classical SOC models, as they lack separation of time scales and any avalanches could appear combined. A practical consequence of this property is that avalanches can no longer be defined as precise events separated by pauses. Instead, a slightly distinct definition of neuronal avalanches, based on a method of thresholding neural activity, had to be employed. Inspired by a previous computational model (Poil et al., 2012), we introduced an activity threshold θ, which artificially included pauses in the excitatory neural activity $a(t) = \sum_{i=0}^{N^{\mathrm{E}}} x_i(t)$. Specifically, a constant background activity θ was subtracted from $a(t)$, for all time steps t, allowing for reasonably frequent pauses and thus reintroducing temporally separated avalanches. θ was initially set to half of the mean network activity, as proposed by Poil et al. (2012), $\langle a(t) \rangle_t = \mu_{\mathrm{IP}} = 0.1$, and was rounded to the nearest integer for simplicity (since $a(t)$ is also constrained to integer values). Each neuronal avalanche was described by two complementary parameters: duration T and size S. An avalanche began when the network activity increased above the threshold θ, where T was the number of subsequent time steps during which the activity remained about such threshold. The size S was defined as the sum of neuronal spikes exceeding the activity threshold at each time step during an avalanche event (see Fig. 4.6B for a graphical example). Formally, for an avalanche starting at t_0, S was given by:

$$S = \sum_{t=t_0}^{t_0+\mathrm{T}} [a(t) - \theta] \tag{4.1}$$

4.2.2 SORN's spontaneous activity shows power-laws

We initially simulated networks of $N^{\mathrm{E}} = 200$ excitatory and $N^{\mathrm{I}} = 40$ inhibitory neurons for a total of 5×10^6 time steps. The neuronal avalanches were measured after the network self-organized into the stable phase. The observed bursts of spiking activity had various sizes and durations, whose distributions could be fit by power-laws for different ranges, up to a size-dependent cut-off point (Fig. 4.7A and 4.7B). For the size, the power-law distribution was a good fit for approximately two orders of magnitude, while the duration is only well fit for approximately one (Fig. 4.7F). The cut-offs observed in the distributions' tails could not be included in any fit, even when considering power-laws with exponential cut-offs ($\propto x^{-\alpha^*} e^{-\beta^*_\alpha t}$), and thus were hypothesized to be the result of a finite size effect. Indeed, with increasing network size the power-law distributions extended over larger ranges (Fig. 4.7C and Fig. 4.7D), while the exponents remained roughly the same (avalanche's duration: $\alpha \approx 1.45$; avalanche's size: $\tau \approx 1.28$). Thus, both for simplicity and for the sake of computational time, we kept the SORN size constant for the remaining simulations of spontaneous activity.

Importantly, we employed maximum likelihood estimators and compared the goodness of fit of power-laws with other possible distributions (Clauset et al. (2009); see also Appendix B). Following the approach by Alstott et al. (2014), we compared the loglikelihood ratio R between power-laws and exponential fits, and between power-laws and stretched exponential fits. The ratio showed that among those distributions, power-laws were better fits for our results (Table 4.1). We refrained from a comparison with more complex distributions with two or more parameters as they might overfit the data, thus leading to an apparent better fit than pure power-laws with only one parameter.

Distribution	size (S)	duration (T)
Exponential ($e^{-\beta t}$)	$R \approx 552$	$R \approx 151$
Stretched exponential (e^{-t^β})	$R \approx 39$	$R \approx 37$

Table 4.1: Goodness of fit (loglikelihood ratio) between power-laws and exponential distributions. A bigger ratio R indicates that a power-law fit is more likely than the compared distribution.

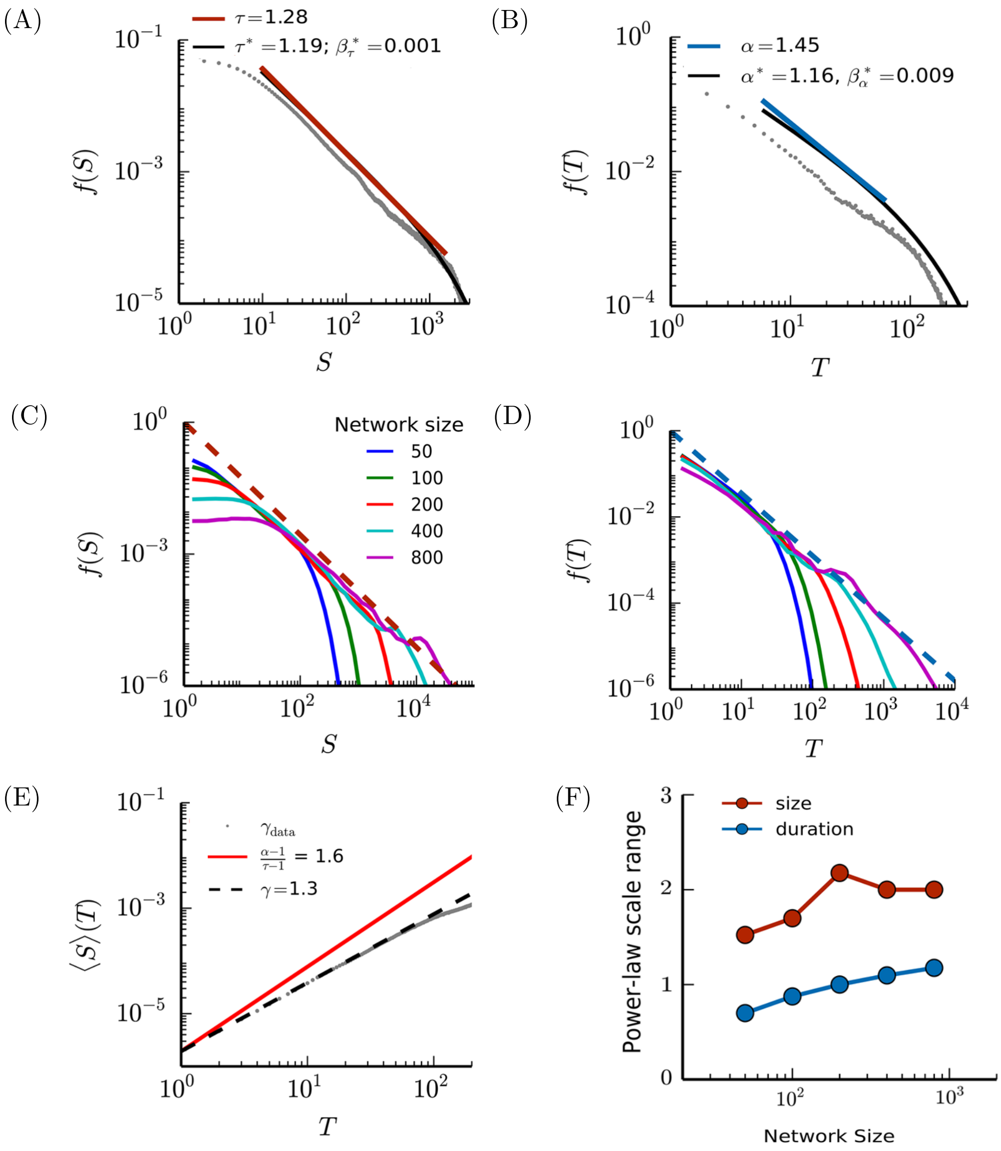


Figure 4.7: **Power-law distributed neuronal avalanches in the $\mathbf{SORN_Z}$'s stable phase.** (A), (B) Normalized distributions of size S and duration T, respectively, for $N^{\mathrm{E}} = 200$. Raw simulation data points are shown in gray. Power-law fits are shown in red/blue, and best power-laws with exponential cut-off fit in black. (C), (D) Scaling of avalanches for networks of increasing size. (E) Avalanche average size as a function of duration, for simulated data (gray) and theoretical prediction (red). The dashed line shows a pure power-law with exponent $\gamma = 1.3$. (F) Power-law scale range, up until the cut-off point, as a function of network size. All distributions show combined results from 50 independent simulations. Figures have been adapted with permission from Del Papa et al. (2017).

The expected theoretical ratio between the power-law scale exponents $(\alpha - 1)/(\tau - 1)$ inferred from the power-law fitting, however, did not match the exponents obtained from the avalanche raw data (Fig. 4.7E), although the average avalanche size did follow a power-law as a function of avalanche duration, with exponent $\gamma_{\text{data}} \approx 1.3$. Recall that, in addition to neuronal avalanches, this ratio is an indicator of critical dynamics in neural networks (Beggs and Timme, 2012). It is worth noting that, although the predictions were not compatible, our numerical exponent γ_{data} agreed with the one calculated directly from experimental data from cortical activity in a previous experimental study (Friedman et al., 2012). This discrepancy might be the result of our thresholding process, a simple consequence of the complex model topology, or even an indication that besides showing neuronal avalanches, the SORN_Z was not near a critical point.

Additionally, we looked at the robustness of the activity threshold and how its choice might have affected our exponents. The activity threshold, which defines the start and end of avalanches, should in principle affect the avalanche distributions since the slope of the power-laws might depend on its choice. Intuitively, small thresholds should increase the avalanches' duration and size while reducing the total number of events. Large thresholds are expected to reduce an avalanche's duration and size while also reducing the number of events. Our results agreed with this intuition, but also showed that the power-law scaling was robust for a range of thresholds, roughly between the 5% and 25% activity percentiles (Fig. 4.8B). Notably, this window contained the previous threshold definition of $\theta = \langle a(t) \rangle_t / 2$ (approximately the 10% activity percentile for a network of size $N^\text{E} = 200$ — see Fig. 4.8A), offering support to our threshold choice. Additionally, as long as the activity threshold remained inside this window, the distributions were better fit by power-laws compared to other single parameter functions (see Table 4.2 for the example loglikelihood ratios between power-laws and exponentials), although their exponents varied. This variation was, however, expected as a result of the activity thresholding process: higher thresholds reduced the relative number of large avalanches (by reducing their size and duration) resulting in steeper curves with larger (in terms of absolute value) exponents. As the attentive reader might perceive, those exponents are close to but still different and slightly smaller than the theoretical predictions for randomly connected networks and critical branching processes that have also been experimentally observed *in-vitro* ($\alpha = 2$ and $\tau = 1.5$; Beggs and Plenz (2003)). This difference may again be due to the fact that SORN models have a complex dynamic topology that differs from random networks after self-organization, or simply an indication of non-critical dynamics even in the presence of some criticality signatures. A mathematical relation between exponents of thresholded avalanche events (without separation of time scales) and exponents from pure SOC systems is, however, not entirely clear (but see the discussion section of this chapter), and any claim about exact critical dynamics in the SORN spontaneous activity remains speculative as long as the origins of these power-laws are not better understood.

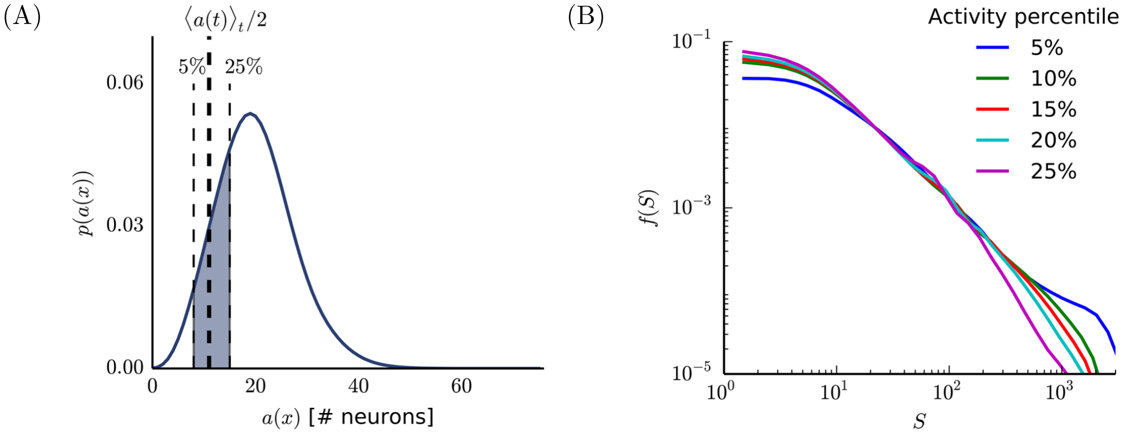


Figure 4.8: **Robustness of choice of activity threshold.** (A) Activity distribution function for the $\mathrm{SORN_Z}$ ($N^{\mathrm{E}} = 200$). The shaded area shows the approximate region where the power-laws hold. The activity peak, set by the target firing rate, is 10% of N^{E}, and the thick dashed line shows half of the average network activity $\langle a(t)\rangle_t$, a common threshold choice. (B) Avalanche size distribution for different activity thresholds θ set as activity percentiles. Although showing different exponents, the power-laws hold for different threshold values (as seen, for example, for θ set at the 5th or 10th percentiles of activity distribution). Curves show combined data from 50 independent simulations. Figures have been reproduced from Del Papa et al. (2017) with permission.

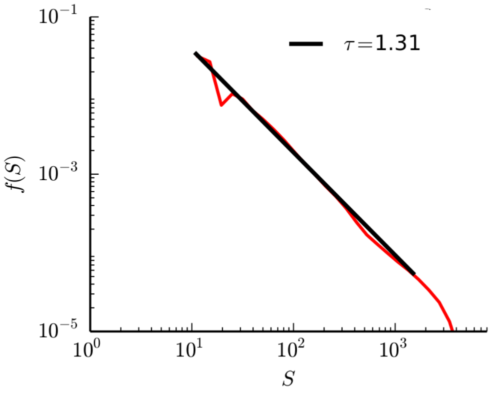
Before further discussing the mechanisms that might originate the power-laws and their relation to a potential critical dynamics, one difference between the $\mathrm{SORN_Z}$ power-laws and experimentally observed ones should be addressed. Independently of the threshold value, $\mathrm{SORN_Z}$ power-laws have a left cut-off for the avalanche size, which is absent in experiments (compare, for ex-

Threshold (θ)	size (S)		duration (T)	
	τ	R_{exp}	α	R_{exp}
5%	1.24674	≈ 836	1.46122	≈ 230
10%	1.31176	≈ 981	1.46809	≈ 284
15%	1.35339	≈ 894	1.52231	≈ 269
20%	1.40378	≈ 772	1.62024	≈ 235
25%	1.52244	≈ 534	1.93752	≈ 191

Table 4.2: Power-law exponents and goodness of fit (loglikelihood ratio between a power-law and an exponential fit; R_{exp}) for different choices of activity thresholds θ ($N^{\mathrm{E}} = 200$).

ample, the raw data in Fig. 4.7A with the distributions in Fig. 4.1A). This unusual cut-off was a consequence of our avalanche size definition (Eq. 4.1). Removing the explicit dependence on θ (i.e., introducing an alternative avalanche definition where the avalanche size was given by $S' = \sum_{t=t_0}^{t_0+\mathrm{T}} a(t)$ instead) modified the left cut-off shape, but did not have a significant effect on the power-law ranges or exponents (Fig. 4.9). In practice, the only difference between our standard definition and this alternative one was a small, proportional increase in all avalanche sizes.

Figure 4.9: **Alternative avalanche definition.** Example of avalanche size distribution (red) and power-law fit (black — exponent $\tau \approx 1.31$) for an alternative avalanche definition (see text). The main effect of removing the explicit dependence of S on θ was seen at the left cut-off, while τ remained largely unaffected. Figure adapted from Del Papa et al. (2017) with permission.



Last, after addressing the issue of threshold dependence, we looked at the effects of our binning process on the distributions. As discussed previously, experimental neuronal avalanches are typically independent of time bin size for SOC models, but not for subsampled or driven systems (Priesemann et al., 2014). Given our definition of avalanche events, we have used a single "unitary" time bin (i.e., one time step) because differently from experiments we could easily look at every single spike and its precise timing. However, in order to fit the power-law distributions, we required a different binning process, relative to the mathematical fit (i.e., the bin size of frequency distributions). For power-laws, the fit of the slope was best done with logarithmic bins, whose width increased in proportion to the variable, reducing fluctuations in the tail of the distributions (Clauset et al., 2009). We tested different logarithmic bin sizes (b_s — see Fig. 4.10) and with the exception of extreme cases, the frequency distributions for different bin sizes virtually overlapped, ruling out any interference of this binning process in our power-law slope fits.

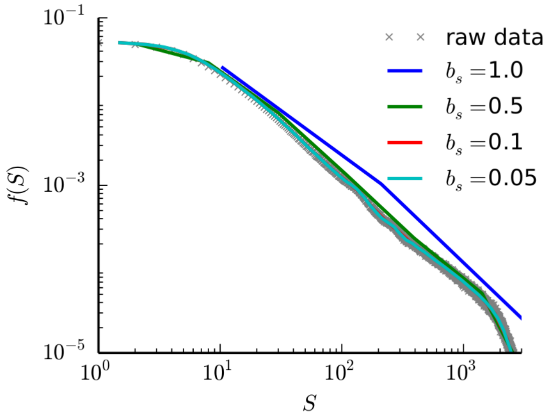


Figure 4.10: **Effects of logarithmic binning on the avalanche distributions.** Varying the logarithmic bin size b_s did not result in significant changes in the power-law ranges, shapes, or exponents. Results for $N^{\mathrm{E}} = 200$, combining data from 50 independent simulations. Figure adapted from Del Papa et al. (2017) with permission.

4.2.3 Criticality signatures are not the result of ongoing plasticity

Having established that the SORN$_{\mathrm{Z}}$ displayed robust power-law distributed neuronal avalanches and that they are not an artifact of our activity thresholding process, we set to investigate what mechanisms might cause them. First, we looked at the role of network plasticity, the SORNs' main components: are they necessary to drive the network into a regime in which the power-laws appear? We compared our results to a SORN model with no plasticity action, which is equivalent to a randomly initialized network. The avalanche distributions observed in the random networks, for both duration and size, did not show power-laws and resembled exponential distributions (Fig. 4.11A, red curves, duration distributions omitted for simplicity). The conclusion was that some plasticity mechanisms, or possibly a combination of them, were necessary for the appearance of criticality signatures.

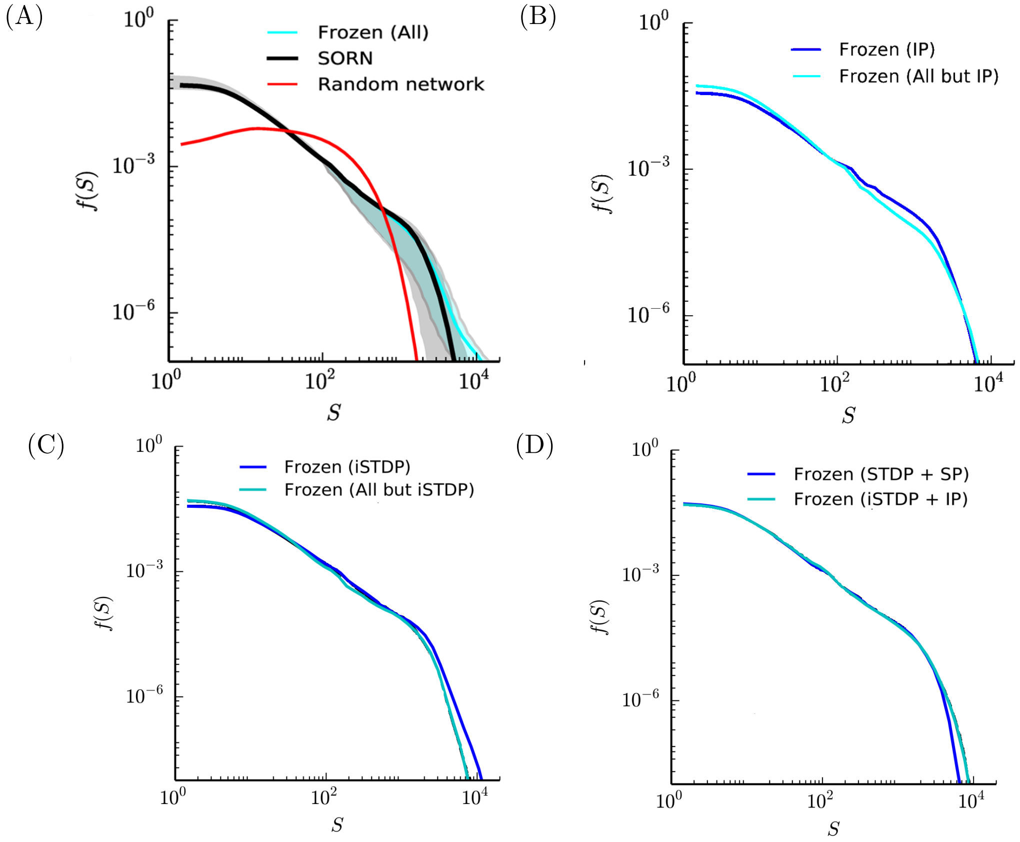


Figure 4.11: **SORN$_Z$ with frozen plasticity.** (A) Distribution of avalanche sizes for the normal SORN$_Z$ (black), random networks without plasticity (red), and a frozen network, with plasticity mechanisms turned off at the stable phase (cyan). Shaded regions show the effects of variations in the activity threshold (5% to 25% percentiles window). (B), (C), (D) Distributions of avalanche sizes for partially frozen networks, with different combinations of plasticity mechanisms turned off at the stable phase. Figures adapted from Del Papa et al. (2017) with permission.

After verifying that the combination of plasticity mechanisms was indeed necessary to drive the network from a randomly initialized state towards a state in which the power-laws appear, we asked whether this result is purely due to the continued action of such mechanisms. If the power-laws appear only when plasticity is active, they could be only a direct result of this ongoing plasticity. If the power-laws held even when all plasticity is turned off (after self-organization), the interpretation that plasticity mechanisms drive the network structure to a special state with criticality signatures would gain more support. We compared, therefore, our previous results with the distributions observed for a frozen network: a network where all plasticity mechanisms where turned off in the stable phase. The $\mathrm{SORN}_{\mathrm{Z}}$ was simulated up until that point, after which the simulation was divided in two: a normal network and a frozen network. We used the same random seed for the membrane noise in both cases so that differences due to randomness were avoided and initialization bias effects could be ruled out. The frozen network resulted in virtually identical power-law distributions for durations (omitted from the figures for simplicity) and sizes (Fig. 4.11A), and the only significant differences were observed in their tails: with frozen plasticity, an increase in the number of large avalanches was observed. This effect could be partly explained by the absence of homeostatic mechanisms that control network activity in the normal, non-frozen network. Likewise, freezing individual or combinations of mechanisms (for example, IP or STDP + SP) did not affect the overall avalanche distributions (see Fig. 4.11B, Fig. 4.11C, and Fig. 4.11D for some examples), indicating that they were not the result of continued action of any particular plasticity rule from the model.

Taken together, these results showed that the $\mathrm{SORN}_{\mathrm{Z}}$'s plasticity mechanisms allowed the network to self-organize into a regime where it showed signatures of criticality. However, the continued action of the plasticity mechanisms was not required for maintaining these criticality signatures, once the network has self-organized.

4.2.4 Noise level contributes to the maintenance of the power-laws

Since a combination of plasticity mechanisms seemed to play a key role in driving the $\mathrm{SORN}_{\mathrm{Z}}$ towards a possible critical regime, our next step was to investigate whether the criticality signatures depended on other model parameters. In particular, we looked at the neuronal membrane noise $\xi(t)$, drawn independently for each neuron as Gaussian noise with $\mu = 0$ mean and variance $\sigma^2 = 0.05$. These parameters are of special interest as they could also be interpreted as a random input level and potentially help with insights not only for the spontaneous activity case but also during an externally driven dynamics.

We found that the avalanche and activity distributions suggested three different regimes depending on the noise level. In the case of high noise levels ($\sigma^2 = 5$), the neurons behaved as if they were statistically independent[11], thus breaking down the power-laws and showing binomial activity centered at the number of neurons expected to fire at each time step (i.e. the mean of the firing rate distribution H_{IP}; Fig. 4.12D). Low noise levels ($\sigma^2 \approx 0$) resulted in a distribution of avalanche sizes resembling a combination of two exponentials, while the activity occasionally died out completely for periods of a few time steps (Fig. 4.12A). A close look at the raster plots of excitatory neuronal activity (Fig. 4.13) also revealed that large bursts of activity only happened at intermediate noise levels ($\sigma^2 \approx 0.05$), while low noise levels resulted mostly in shorter bursts and high noise levels resulted in Poisson-like activity. Therefore, we concluded that, together with the plasticity mechanisms, the noise level determined the network dynamical regime. Taken together with the drastic changes in activity distribution for each regime (Fig. 4.12A), these results supported the existence of a phase transition point, with the noise level (or its variance) acting as a control parameter.

In order to further investigate the contribution of the noise level to the maintenance of the criticality signatures, we tested if other types of noise could have a similar effect on the network's dynamical regime, and how diffuse this noise needed to be in order to allow for the appearance of the power-laws. First, we switched from Gaussian noise to random spikes: each neuron received input surpassing its threshold with a small probability of spiking at each time step p_s. Using p_s as a control parameter in the same way as the Gaussian noise variance, we could reproduce all of the previous findings: three different distribution types including a transition window in which the power-law distributions of neuronal avalanches appeared (Fig. 4.12B and Fig. 4.12E). Second, we observed that limiting the noise action to a subset

---

[11] This result is, naturally, expected for a network of randomly spiking neurons, independently of its topology. Such random firing regime can also be interpreted as a sanity check of our implementation of the model.

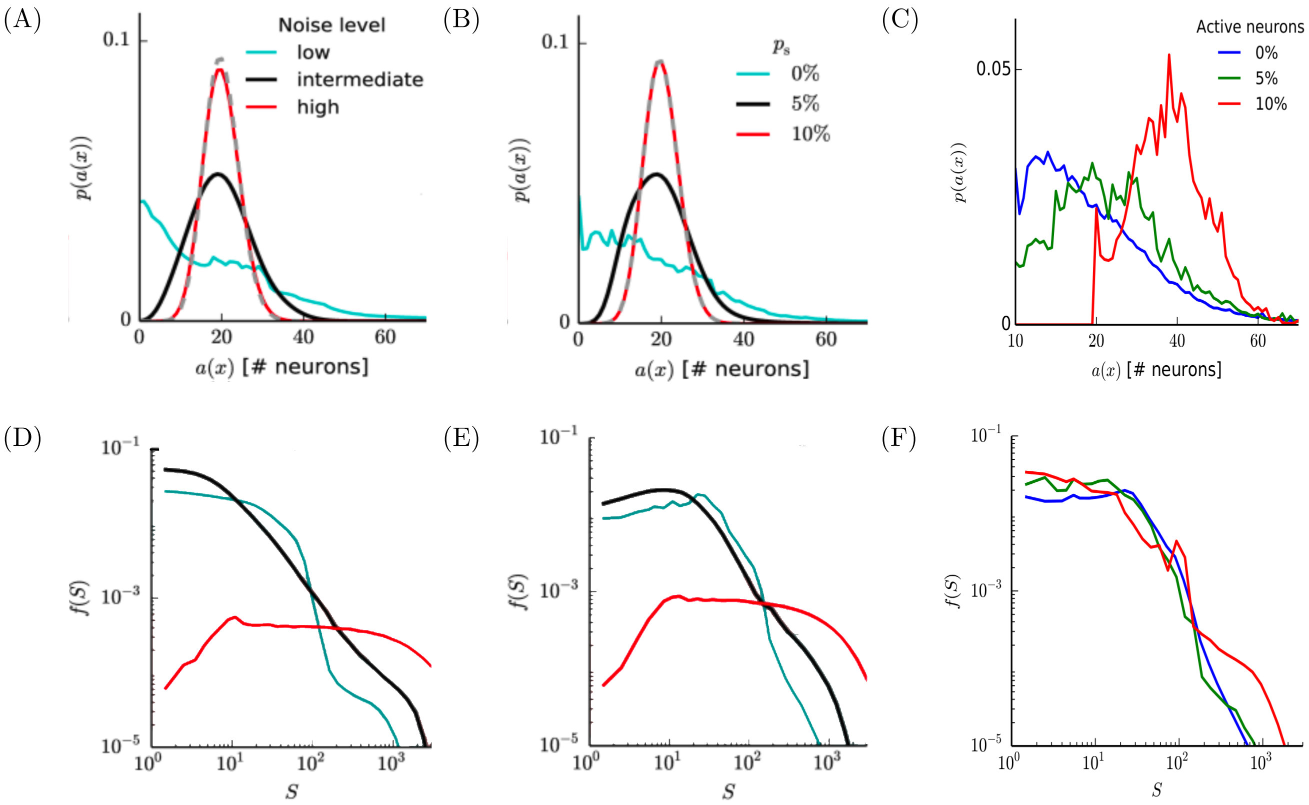


Figure 4.12: **Noise level gives rise to criticality signatures.** Top row: activity distributions for distinct noise sources and levels. Bottom row: distributions of avalanche sizes for the same noise forms. (A), (D) Gaussian noise levels: low ($\sigma^2 = 0.005$), intermediate ($\sigma^2 = 0.05$) and high ($\sigma^2 = 5$). Very weak or strong noise levels break down the power-laws, suggesting non-critical regimes. (B), (E) Random spike noise source (see text), with different probabilities of random spiking p_s. Results are strikingly similar to Gaussian noise, with three distinct regimes. Black curves show the standard SORN$_Z$ parameters that give rise to power-law distributed neuronal avalanches. Dashed gray lines show a binomial distribution, for comparison. (C), (F) Noise limited to randomly chosen subsets of excitatory neurons shows no power-laws or signs of criticality. Percentages indicate the number of excitatory units receiving suprathreshold input at each time step. Figures were adapted from Del Papa et al. (2017) with permission.

of units[12], while keeping all plasticity mechanisms on, abolished power-laws completely (Fig. 4.12C and Fig. 4.12F). Different subset sizes were compared (10%, 5%, or 0% of excitatory units were continuously active), while the activity threshold was set again to half of the mean network activity, but now excluding the subset of continuously active units. We concluded that the observed criticality signatures required not only a specific noise level, suggesting a control parameter for the network, but also a relatively even distribution of noise across the network units. Additionally, the intermediate noise level resulted in a phase transition state between two very different dynamics, again suggesting a critical or near critical point.

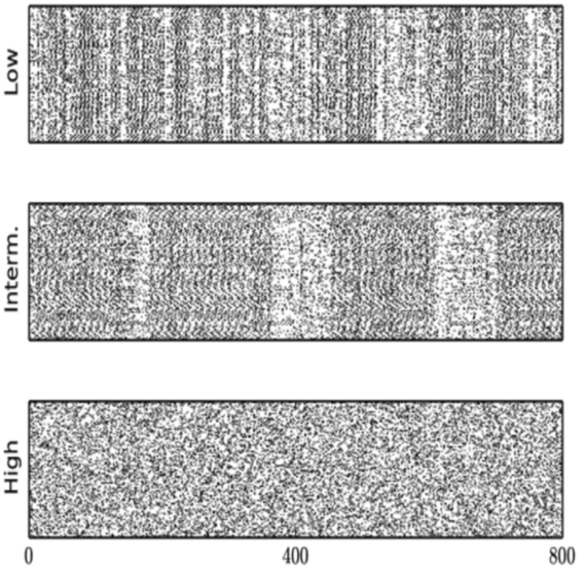


Figure 4.13: **Raster plots for distinct noise levels.** Typical raster plots of excitatory activity at low ($\sigma^2 = 0.005$), intermediate ($\sigma^2 = 0.05$), and high ($\sigma^2 = 5$) Gaussian noise levels. Bursts of activity of varied sizes only appear in the intermediate case and are short-lived (low noise levels) or nonexistent (high noise levels). Figure adapted from Del Papa et al. (2017) with permission.

Finally, if the neuronal membrane noise level was capable of deviating the network away from a regime in which criticality signatures appear, but the plasticity mechanisms seemed to drive the network dynamics back towards this state, a few new questions arise. What happens under external input that differs from noise? Can the power-laws be maintained independently of the input intensity, or does the $\mathrm{SORN_Z}$ reproduce subcritical dynamics as observed *in-vivo*?

---

[12] Note that we keep here the *noise* nomenclature for consistency only, since suprathreshold input limited to a fixed subset of units is equivalent to deterministic external input.

4.3 External input: readaptation and learning

Neural circuits in awake, behaving animals typically receive a large number of different external inputs. For the critical brain hypothesis to hold true, adaptation mechanisms must tune neural circuits towards criticality under various external input regimes. We have seen so far that this adaptation is not only possible in computational models, but criticality also improves various input encoding capacities. Additionally, experimental evidence has shown that the onset of external input breaks down criticality signatures in the form of power-laws (Shew et al., 2015; Clawson et al., 2017). Specifically, both studies showed that cortical *ex-vivo* activity measured in the turtle brain is not critical immediately after the onset of strong external input, but critical dynamics quickly reappears due to the system's readaptation mechanisms.

What could be, therefore, the role of this readaptation? As criticality improves essential encoding capabilities, a critical regime is also extremely beneficial for the learning capacities of various types of networks, and one can hypothesize that the same mechanisms that drive a network towards criticality are also responsible for its learning abilities. Potentially, both processes might be intrinsically connected in neural circuits by self-organization mechanisms. In this section, we study the possible link between a potentially critical dynamics (or at least, dynamics displaying criticality signatures) and learning by looking at distinct input conditions: random and structured. Both input types affect the model's internal dynamics and can, in principle, drive it towards different states, but only the latter has useful patterns that can be learned by a network model. We further analyze the effects of different input patterns on neuronal avalanches and compare the $\mathrm{SORN_Z}$ activity with experimental externally driven networks, proposing that a combination of plasticity mechanisms is responsible for the quick readaptation after external input onset.

4.3.1 Network readaptation after external input onset

In a first experiment, random external input (sequence of length $U_\mathrm{L} \to \infty$ and alphabet size $U_\mathrm{A} = 10$, see section 3.2.3 for details on the definiton of those values) was presented to the network after it reached the stable phase, and distributions of avalanche size and duration were measured first immediately after its onset, including avalanche events starting in the first 10 time steps after the stimulus, and second after readaptation due to plasticity, with all plasticity mechanisms on (2×10^6 time steps). The number of time steps defining the size of the transient period was chosen based on what has been commonly chosen in experiments, roughly hundreds of milliseconds (Shew et al. (2015); recall that each time step in the SORN models corresponds to the typical time scale of STDP, around 10 to 20 ms). The external input consisted of a random sequence of symbols, where each symbol provided strong input ($u_i^{\mathrm{Ext}}(t) = 10^5$) to a subset of N^{U} excitatory neurons (see Chapter 3 for further details and the exact update equations). Importantly, the activity threshold θ was kept the same for both time periods.

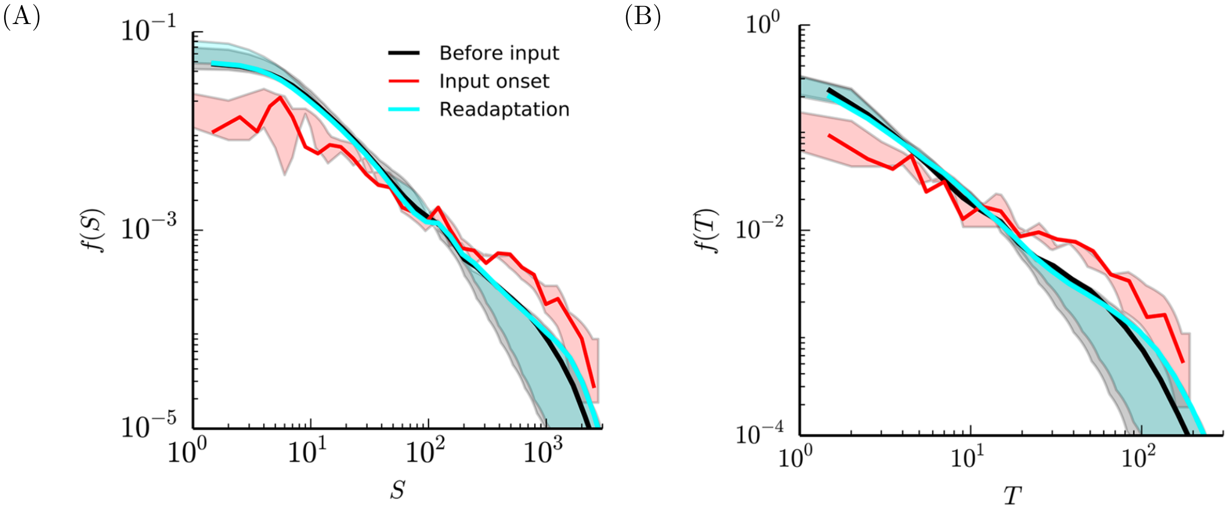


Figure 4.14: **SORN$_{\mathrm{Z}}$ readaptation after random external input onset**. Size (A) and duration (B) distributions of avalanche events before input onset (black), during a short transient readaptation period (red) and after readaptation (cyan). During the transient readaptation period, power-laws are modified and bigger avalanches become relatively more frequent. Before input and Readaptation curves show combined data from 50 independent simulations, while Input onset curves show data from 250 input trials. Shaded areas represent the area between the curves obtained for activity thresholds at the 5th and the 25th percentiles. Figures reproduced from Del Papa et al. (2017) with permission.

In agreement with the *ex-vivo* recordings (Shew et al., 2015), external input onset resulted in flatter power-laws (Fig. 4.14, red curve). As in the experimental recordings, we also observed network readaptation towards the regime in which power-laws appeared, after a transient period (cyan curve). Furthermore, the flatter power-laws and the subsequent readaptation appeared even under weaker external inputs ($u_i^{\mathrm{Ext}}(t) \approx 1$, in contrast to the default strong input). This finding supported the hypothesis that plasticity mechanisms were responsible for driving the $\mathrm{SORN_Z}$ towards the appearance of criticality signatures, even after transient changes due to external stimulation. Interestingly, this result suggested that neuronal avalanches are, indeed, input dependent, and the same network might display them or not depending on the input intensity and structure. This result could help to clarify and reconcile *in-vitro* observed criticality signatures and their absence in spiking activity *in-vivo*, as the external input is one of the key differences between those experimental setups.

4.3.2 Absence of criticality signatures under input of learning tasks

In contrast to random external input, structured input is used in spatio-temporal learning tasks and is commonly associated with more realistic sensory input in behaving neural circuits (Lazar et al., 2009; Hartmann et al., 2015). We studied two simple learning tasks that required the $\mathrm{SORN_Z}$ to encode temporal information about its input: a Counting Task (Lazar et al., 2009) and a Sequence Task. We discuss both tasks separately below.

Counting Task

The Counting Task was inspired by one of the learning tasks in which the original $\mathrm{SORN_L}$ model was shown to outperform static reservoirs (Lazar et al., 2009). It consisted of randomly alternating sequences of symbols of the form "ABB...BBC" and "DEE...EEF", with n middle, repeating symbols. Each symbol provided extra input, $u_i^{\mathrm{Ext}}(t) = 1$, for a randomly selected but fixed subset of excitatory neurons N^{U} at the time step in which it was presented to the network. Differently from the former random external input experiment, these sequences were presented during the whole simulation, one symbol per time step. The model was trained with plasticity (all plasticity mechanisms active) for $T_{\mathrm{plast}} = 5 \times 10^4$ time steps, and the performance was evaluated by training a readout layer to predict the *next* symbol (input at $t+1$) based on the network internal state (i.e., the recurrent activity without the external input term, u_i^{Ext}) at time t. The readout was trained for $T_{\mathrm{train}} = 5 \times 10^3$ time steps and evaluated for another $T_{\mathrm{test}} = 5 \times 10^3$. The final performance was the percent of correct predictions ignoring the first symbol of each sequence, as both sequences could appear with the same probability. Generally speaking, the model would have to learn how to "count" the number of middle symbols to correctly predict the final one (thus the name of the task).

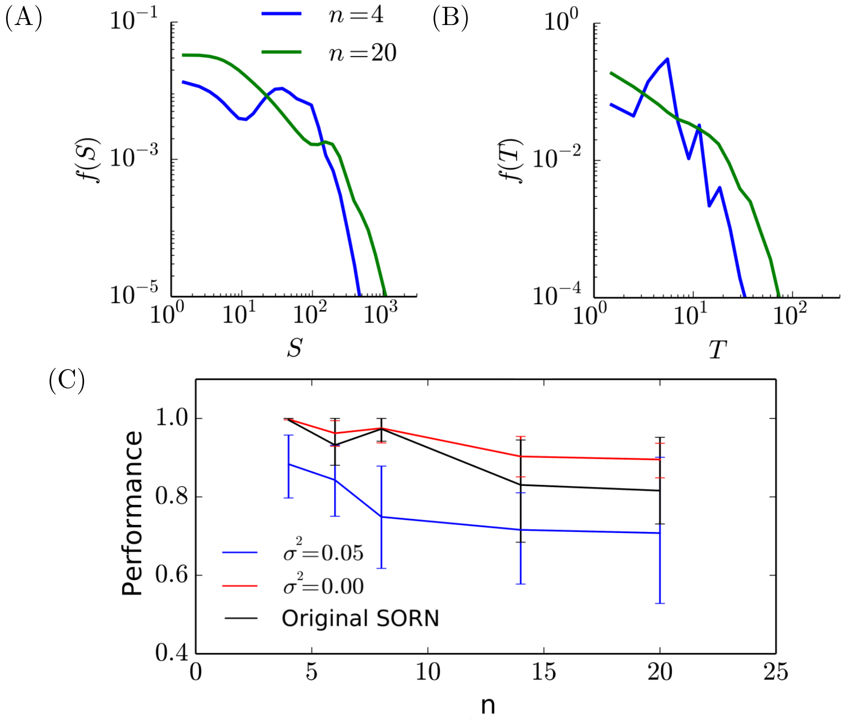


Figure 4.15: **Learning a Counting Task.** (A), (B) Size and duration distributions, respectively, during the Counting Task for different input sequence lengths with n middle symbols, in the presence of membrane noise ($\sigma^2 = 0.05$). (C) $\mathrm{SORN_Z}$ performance as a function of the sequence size n, for two different membrane noise levels. Original SORN refers to the $\mathrm{SORN_L}$ model (without iSTDP, SP, and membrane noise; Lazar et al. (2009)). Curves show the average of 50 independent simulations and error bars show the 5% to 95% percentile interval. Figures adapted from Del Papa et al. (2017) with permission.

We measured the avalanche distributions for duration and size after the active plasticity period and verified that the power-laws did not appear in this case, independently of n (Fig. 4.15), although the distributions appeared smoother and visually more similar to power-laws for longer sequences (larger n). This finding suggested that structured input did not allow for the appearance of the power-laws. In this case, in contrast to random external input, our plasticity mechanisms could not drive the network towards the supposed critical regime and counteract the input effect. Furthermore, we measured the performance of the $\mathrm{SORN_Z}$ in the task and found that this model was capable of maintaining a performance higher than 90% when the membrane noise was removed ($\sigma^2 = 0$; Fig. 4.15C), which is consistent with the results obtained in the original SORN model for the same task (Lazar et al., 2009). With the addition of membrane noise ($\sigma^2 = 0.05$), however, we observed decay in the overall performance, particularly for long sequences.

Sequence Task

The Sequence Task consisted of a different form of external input with longer sequences and a larger number of symbols compared to the Counting Task. At the beginning of each simulation, we defined a random sequence of size L, which would become the discrete input to be subsequently repeated indefinitely. The training procedure was again similar: during $T_{\mathrm{plast}} = 5 \times 10^4$ time steps, the model was subject to input while all plasticity mechanisms were active. Each input symbol provided additional input to a subset of excitatory neurons in the same manner as the Counting Task. Later, a readout layer was trained for $T_{\mathrm{train}} = 5 \times 10^3$ to predict the *next* element of the sequence (at $t+1$), again based on the model's recurrent activity (at t). Now, however, as the same sequence was repeated, there was no need to exclude the last symbol from the performance evaluation, which was simply the percent of correct predictions over the last $T_{\mathrm{test}} = 5 \times 10^3$ time steps.

The action of the plasticity mechanisms abolished the criticality signatures under structured input (Fig. 4.16), but, as observed in the Counting Task, longer sequences showed smoother curves. Those are the same plasticity mechanisms, however, that improved performance on a sequence learning task in the original $\mathrm{SORN_L}$, as well as in the $\mathrm{SORN_Z}$, compared to a randomly initialized reservoir (RR; Fig. 4.16C). Interestingly, the performance in the $\mathrm{SORN_Z}$ was better under low membrane noise than under medium noise and decreased to chance level for high noise (not shown). This, on the one

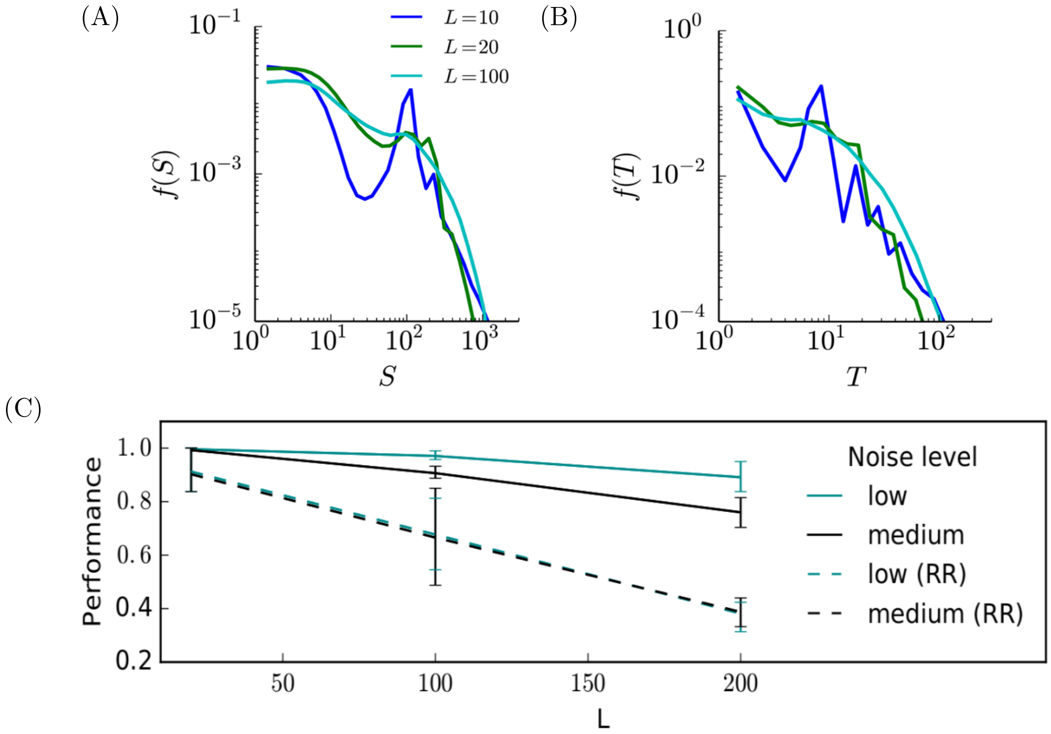


Figure 4.16: **Learning a Sequence Task.** (A), (B) Size and duration distributions, respectively, during the Sequence Task for different input sequence lengths L, in the presence of membrane noise ($\sigma^2 = 0.05$). (C) SORN$_\mathrm{Z}$ performance as a function of the sequence size L, for two different membrane noise levels (low → $\sigma^2 = 0.005$; medium → $\sigma^2 = 0.05$). For the SORN$_\mathrm{Z}$, performance depends on the membrane noise level, while this is not the case for random reservoirs (RR). Curves show averages of 50 simulations and error bars indicate the 25% to 75% percentiles interval. Figures adapted from Del Papa et al. (2017) with permission.

hand, was to be expected, as membrane noise masked the input sequences. On the other hand, such a result may appear surprising at first sight, as the SORN$_\mathrm{Z}$ showed signatures of criticality under intermediate, but not low or high noise levels.

In summary, what do these tasks mean for the relation between criticality and information processing? Surprisingly, during such tasks, the repeating structure of the input was enough to destroy the power-law distributions in the $\mathrm{SORN_Z}$ activity. Yet, longer sequences resulted in smoother distributions, qualitatively closer to power-laws, as long repeating input sequences resembled random input, suggesting that additionally to plasticity, the structure of the external drive fundamentally controlled the model's dynamical regime. Regarding the model's performance, a few conclusions could be drawn. First, one may conclude that performance is maximized at a state that does not show power-laws. However, as none of the tested conditions showed power-laws, it is still conceivable that there may exist a state with power-law scaling and even better performance. Second, the task is fairly simple, as it predicted the pattern at $t+1$ from the activity at t. Maximal performance at this task may not require critical dynamics. Note that criticality maximizes certain input encoding properties, such as susceptibility, correlation length and time, and pattern diversity. Maximization of these properties fosters performance in tasks that rely on them, e.g. tasks that require maintaining information about past input in their activity for a long time (random reservoir properties). However, for simple tasks as the one used here, fast forgetting might be of advantage (Boedecker et al., 2012). Hence, the higher performance in the simple task under low membrane noise was expected.

In the next chapter, we continue to directly tackle the learning properties of the SORN models, switching focus to their memory capacity, the ability of a network to store information about past inputs. Before this small change of topics, however, we summarize our current conclusions about the possibility of a critical regime in the SORN dynamics, given all criticality signatures we have measured so far, and its effects to the variations in performance on these simple learning tasks.

4.4 More than one critical point?

From the brain's perspective, critical dynamics is highly desirable due to its many functional benefits (Shew and Plenz, 2013). Specifically, networks tuned to criticality showed maximal dynamical range (Kinouchi and Copelli, 2006), maximal information capacity (Shew et al., 2011), maximal number of metastable states (Haldeman and Beggs, 2005), and even maximal complexity in neural systems (Timme et al., 2016). A similar concept, known as computation at the edge-of-chaos (i.e., at a phase transition point separating non-chaotic from chaotic dynamics) is known to increase performance in classification tasks (Legenstein and Maass, 2007) and maximize information transfer and storage (Boedecker et al., 2012) in recurrent neural networks, allowing them to perform complex computational tasks due to an increased fading memory capacity (Bertschinger and Natschläger, 2004). Edge-of-chaos, nonetheless, is also referred to as criticality because chaotic dynamics occur near transition states, which potentially means that different studies use different definitions for criticality and critical points. On the one hand, criticality in the sense of neuronal avalanches, as we have been discussing so far, is typically measured via control and order parameters, whenever possible, or via indirect signatures for more complex models without separation of time scales. These signatures, as we discussed in this chapter, include the power-laws for distributions of bursts of activity, mathematical relations between the distributions' slopes, ability to tune the system with a control parameter (as the excitation and inhibition balance), among others. On the other hand, edge-of-chaos criticality is measured via perturbation analysis and Lyapunov exponents (Kanders and Stoop, 2016), which have, up to this point in time, being largely unexplored in experiments due to experimental constraints (i.e., systematically controlling and tracking perturbations either *in-vitro* or *in-vivo* is particularly challenging and/or infeasible with current techniques). Although those two definitions are sometimes assumed to refer to the same dynamical state and commonly considered to coexist, they might in principle refer to distinct phenomena, which increases the problem of comparing different models displaying any indication of criticality.

In fact, it has been recently proposed that both "types" of criticality may occur independently in neural networks (Dahmen et al., 2016, 2017; Kanders et al., 2017a). Furthermore, when considering neural systems without separation of time scales, different "routes" towards a critical (avalanche) dynamics have been observed (Taylor et al., 2013; Hartley et al., 2014), suggesting that the lack of co-occurrence between neuronal avalanches as criticality signatures and a critical point might imply that these refer to distinct dynamical states. Such hypothesis has at least some theoretical backing, as one computational model has suggested that biologically inspired systems give rise to avalanche criticality, but at a point where edge-of-chaos criticality fails to co-occur ((Kanders et al., 2017a,b); see Fig. 4.17 for an example of this model's results). This model, however, relied on one important detail: a single "leader neuron" artificially tuned to fire in a chaotic manner, which might lead the system to a state with the coexistence of critical and non-critical dynamics. This combination of two dynamics has been proposed to yield optimal population coding in a slightly different class of models (Gollo, 2017), and arguments on the benefits of smaller, critical subsets of bigger systems have been proposed. These benefits, however, might be specific to some model architectures, as a general link between edge-of-chaos criticality and neuronal avalanches is still missing. Thus, a deeper understanding of the self-organization mechanisms that lead to a supposed avalanche critical regime is necessary in order to describe how useful information processing capacities, typically associated with edge-of-chaos dynamics, might also arise in those cases.

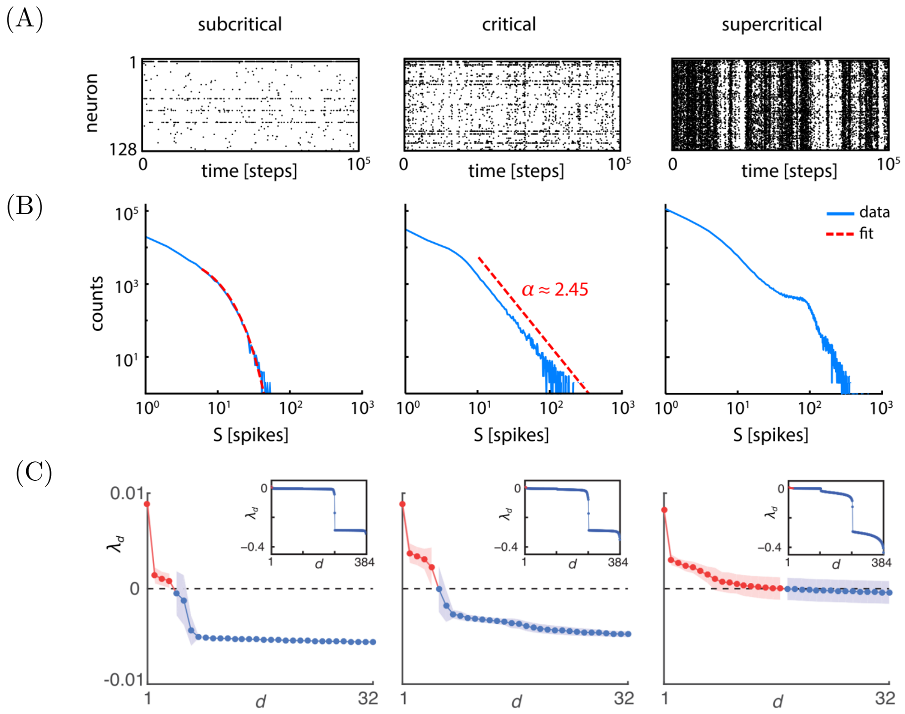


Figure 4.17: **(Kanders et al., 2017a): Lack of co-occurrence between avalanche and edge-of-chaos criticality.** (A) Raster plots for the three dynamical regimes, from subcritical to supercritical, achieved by scaling the weights of a biologically inspired network. (B) Distributions of avalanche sizes. Power-laws appear in a phase transition state, defining criticality in the model. (C) Lyapunov spectra λ_d for each dynamical regime, for the first d exponents (insets show the full spectra). All regimes show positive exponents (red), revealing supercritical (edge-of-chaos) dynamics. Figures reproduced from Kanders et al. (2017a) with permission.

What does the difference between avalanche criticality and edge-of-chaos criticality mean for our results on the SORN models? We have investigated the occurrence of criticality signatures and referred to self-organization towards criticality a number of times. Given the nature of our results, aiming to model experimentally observed phenomena, we have employed the avalanche definition of criticality and have not analyzed the values of Lyapunov exponents in response to perturbations. In fact, such analysis has already been done for the $\mathrm{SORN_L}$ (Lazar et al., 2009) and revealed a slightly subcritical state via an approximation of small perturbations employing the discrete Hamming distance, under the input of the Counting Task. In the $\mathrm{SORN_Z}$, we have observed that criticality signatures also break down under structured input of this particular learning task, raising the possibility that the model is indeed not in a critical state, both in the avalanche and edge-of-chaos sense, when performing learning tasks. Interestingly, this scenario would correspond to *in-vivo* spike activity, which also does not show criticality signatures (Priesemann et al., 2014), but arguably must excel at information processing. In the spontaneous activity case, nonetheless, our observation of power-law distributed bursts of activity did not reveal if the model operates at the edge-of-chaos, and for now, we limit our conclusions to the avalanche definition of criticality (but see Chapter 5 for an additional argument based on the SORN's fading memory capacity).

4.5 Discussion

Criticality is a central concept connecting microscopic and macroscopic levels of a complex system and frequently leads to key insights about the behavior of large systems composed of smaller, similar units (Hesse and Gross, 2014). The hypothesis of criticality in the brain as discussed here is largely based on experimental measurements of power-law distributed neuronal avalanches. This hypothesis is still controversial in the neuroscience community, in particular, because power-law distributions can be generated by a number of other mechanisms but criticality (Touboul and Destexhe, 2010) and thus are not sufficient to prove that a system is critical. For that reason, our neuronal avalanche analysis alone does not prove that the $\mathrm{SORN_Z}$ self-organizes towards a critical state. Instead, we first highlight that the combination of plasticity mechanisms in the model is sufficient to produce the same criticality signatures typically observed in *in-vitro* experiments, independently of the question whether these systems are critical or not.

The measured exponents for duration and size, $\alpha \approx 1.45$ and $\tau \approx 1.28$, were both smaller than those expected for random-neighbor networks (2 and 1.5, respectively), potentially reflecting the complex topology emerging after the $\mathrm{SORN_Z}$ self-organization (or, this discrepancy might be purely a result of the activity thresholding process, which has been proposed to alter those exponents (Font-Clos et al., 2015)). The power-laws typically spanned one or two orders of magnitude for the durations and sizes, respectively, which is comparable to experimentally observed data. Before and after the power-law interval, the size distribution often showed cutoffs. While the right cutoff typically arises from finite size effects (Privman, 1990), the left cutoff is not characteristic for classical critical systems such as a branching network (Harris, 2002), possibly being the result of our avalanche definition based on thresholding the network activity. However, left cutoffs have been observed for neural avalanche distributions in the cortex (e.g., Priesemann et al. (2014); Shew et al. (2015)). Therefore, the neuronal avalanche distributions we observed seemed to be indeed compatible with the experimental ones we described in the first half of this chapter.

Our results also proposed that the combination of biologically inspired Hebbian and homeostatic plasticity mechanisms in the $\mathrm{SORN_Z}$ was responsible for driving the network towards a state in which power-law distributed neuronal avalanches appeared, even though such plasticity action was not required for the maintenance of this state in the case of spontaneous activity. The power-law distributions of avalanche durations and sizes in the $\mathrm{SORN_Z}$'s spontaneous activity replicated a widely observed phenomenon from cultured cortical networks (Beggs and Plenz, 2003; Tetzlaff et al., 2010; Friedman et al., 2012) to awake animals (Petermann et al., 2009; Hahn et al., 2010; Priesemann et al., 2014). Notably, the network also reproduced the short transient period with bigger and longer neuronal avalanches and subsequent readaptation after external input onset which has been observed in the turtle visual cortex (Shew et al., 2015; Clawson et al., 2017). Additionally, the power-laws also required a suitable intermediate level of membrane noise to occur in the model's spontaneous activity. This finding suggested that, in the cortex, the strength of input received from other connected areas could act as a control parameter maintaining the local network circuit in a reverberating state, thus within the subcritical regime (Zierenberg et al., 2018). Therefore, we highlight the role of self-organization in driving a network towards a regime in which criticality signatures appear, but suggest that this regime is only achievable under particular input conditions. If the brain is indeed critical, a remaining question is whether it self-organizes to criticality as a single system or as a collection of many subsystems, as critical dynamical behaviors may emerge even when the underlying dynamical processes are not critical (Friedman and Landsberg, 2013). While computational models consider predominantly homogeneous networks, the brain is composed of a range of diverse subsystems that potentially self-organize independently (Hesse and Gross, 2014). A better understanding of this question could strengthen the link between criticality and its medical, and even technological (Srinivasa et al., 2015), implications.

Previous studies have already identified plasticity mechanisms that tune a network to criticality. For example, networks of spiking neurons with STDP (Meisel and Gross, 2009; Rubinov et al., 2011) showed critical dynamics, and the earliest example of self-organization towards criticality in plastic neural networks is probably the network by Levina et al. (2007), who made use of dynamical synapses in a network of integrate-and-fire neurons. Furthermore, it is known that networks without plasticity can be fine-tuned to a critical state, where they show favorable information processing properties, both in deterministic (Bertschinger and Natschläger, 2004; de Arcangelis et al., 2006; Boedecker et al., 2012) and stochastic (Haldeman and Beggs, 2005; Shew et al., 2011; Poil et al., 2012) systems. Those models are very important to describe the properties of a network already in a critical state. Beyond those results, here we have shown, for the first time, criticality signatures arising in a network model initially designed for sequence learning and cortical dynamics modeling, via a combination of Hebbian and homeostatic plasticity mechanisms.

Linking criticality or the deviation from it to performance in a particular learning task in experiments is particularly challenging, and theoretical work is crucial for their systematic understanding. Experimentally, a large body of work focuses on testing whether recorded neural activity *in-vivo* or *in-vitro* complies with the criticality hypothesis in showing avalanche distributions without a direct link to function. The challenges when investigating learning tasks are twofold. First, tasks typically come with altered input for each condition. The input makes it very difficult to disentangle whether an observed difference is indeed caused by a deviation from criticality or any general state change, whether it is induced by transiently changing (non-stationary) input without underlying state change, or a combination of both. Developing approaches to disentangle the two scenarios is an important future challenge.

The second challenge is that avalanche analysis requires tens of minutes, or even hours, of recordings to be able to detect differences. This is because an avalanche distribution that extends over two or more orders of magnitude comprises thousands of avalanches for sufficient statistics. Assuming a rate of one avalanche per second, the analysis requires at least 1000 seconds (or around 20 minutes of recording — per condition). A fine temporal resolution in state change is, therefore, difficult and data-costly. Current studies focus on estimators in order to quantify the distance of a particular system to criticality (Wilting and Priesemann, 2016), what may in the future enable a larger number of experimental insights linking dynamical state, criticality, and learning task processing.

Finally, the contrast between the presence of power-law distributed neuronal avalanches in the model's spontaneous activity and their absence under structured input also suggests an analogy between *in-vivo* and *in-vitro* activity in the brain. We have shown that the same plasticity mechanisms might result in the occurrence or absence of power-law distributions under different input conditions. As such distributions indicate avalanche criticality, our results stand in agreement with the development of criticality signatures in neural networks *in-vitro* and with the non-critical dynamics observed in spike avalanches *in-vivo*. Such input driven adaptation may be favorable for the neural circuits in behaving animals, as it allows them to take advantage of improved learning abilities and the computational advantages of criticality while avoiding unstable supercritical regimes observed during epileptic seizures. Thus, by keeping neural activity at a healthy, non-epileptic level, biological plasticity mechanisms might play an essential role in tuning the system towards and also away from criticality when required by various input conditions.

Chapter 5

Learning with plasticity: from random sequences to sentence generation

> (...) for memory tasks, however, the input consisted of nonrepeating random sequences and the performance was evaluated on past time steps at each time step, the performance was defined as the normalized number of correct classifications of input symbols received time steps in the past. again, we emphasize that both capacities refer to different phenomena the learning capacity measures the encoding of temporal patterns from the input sequence, we took advantage of a modified sequence learning tasks, which was already employed in the last chapter as an example of structured external input. now, we have additionally shown that plasticity can also improve the fading memory capacity. (...)
>
> SORN*

Having explored the experimental sings of critical dynamics and their emergence after self-organization in recurrent networks, we now switch focus

---

*Quote text autonomously generated by a SORN model with $N^{\mathrm{E}} = 10\,000$, trained for $T_{\mathrm{plast}} = 500\,000$ time steps with the text of the two first sections of this chapter (5.1 and 5.2, excluding figure captions, references, and mathematical symbols). For further details, see section 5.4.2. The code is provided at `https://github.com/delpapa/SORN_V2`.

to another interesting function of general random reservoirs — their memory capacity. Generally speaking, a model's memory is intrinsically connected to its overall learning capacity, allowing it to easily recall recent inputs and thus learn from them. This capacity is also present in biological neural circuits and can roughly be considered as the brain's working memory. Curiously, the theoretical fading memory capacity has been shown to be greatly increased in random networks at the edge-of-chaos, which raises the possibility of further improvements due to plasticity action, in a similar manner than other information processing abilities near critical regimes. We begin this chapter by investigating the co-occurrence of neuronal avalanches and fading memory improvements resulting from biologically inspired self-organization, again by employing self-organizing recurrent neural networks (SORNs). The goal of this analysis is twofold: not only we observe how plasticity and near critical dynamics are able to improve learning and memory abilities, but also we aim to better understand how both definitions of criticality (edge-of-chaos and avalanches) may meet as a result of self-organization. We continue by discussing possible applications of this memory capacity to spatio-temporal tasks and finally investigate the performance of SORNs in a few simple grammar and language learning tasks. For the latter, we take a turn in the direction of machine learning and compare SORNs to simple deep recurrent network architectures. Finally, we finish by taking advantage of their spatio-temporal learning abilities near criticality and propose their use as simple generative models.

5.1 Criticality in reservoir computing

The $\mathrm{SORN_L}$ model (Lazar et al., 2009) was initially proposed as an improved reservoir with time-varying connections and firing thresholds, outperforming classic static reservoirs at various spatio-temporal tasks. Such improvement in performance was a consequence of the plasticity driven dynamics, which was likely maintained near criticality both in its spontaneous activity and under external input. Nonetheless, static reservoirs, and more generally sparsely connected recurrent neural networks, already have their own merits on spatio-temporal learning tasks. Differently than most feedforward models, reservoirs and recurrent networks aim to solve spatio-temporal pattern recognition tasks since they are suited for dynamic data processing instead of only regression and classification. In other words, recurrent neural networks and reservoirs approximate dynamical systems, while feedforward networks approximate functions (Lukoševičius and Jaeger, 2009). Generally, static reservoirs first map inputs onto high-dimensional spaces, from which linear supervised readout layers are able to extract patterns and only then perform regression or classification. This already shows an important advantage compared to dense recurrent neural networks trained with gradient descent methods: reservoirs typically learn faster, avoiding the computational cost of performing backpropagation at each time step. Additionally, due to their relative simplicity and the lack of adaptive weights, reservoirs are more suitable for hardware implementation, for which various physical systems could potentially be used (Paquot et al., 2012; Tanaka et al., 2018).

Given the performance-driven deep learning "revolution" in recent years, computation with reservoirs has become comparatively less popular as a framework for applied artificial neural networks. With the increase in general computation capacity, combined with cheaper parallel computing on GPUs, backpropagation became feasible in reasonable computational time for increasingly complex architectures and deeper networks. In the case of recurrent neural networks, the current popular models are based on Long Short-Term Memory cells (LSTM; Hochreiter and Schmidhuber (1997)) or Gated Recurrent Units (GRU; Cho et al. (2014)), which differently from static reservoirs have adaptive, dense weights updated via variations of backpropagation through time (Werbos, 1988)[1]. Although effective in terms of model performance and generative tasks, these weight update rules suffer from the inevitable shortcomings of their own complexity: the huge number of variables and architecture dependent hyperparameters eventually led to empirical optimizations in spite of a broader understanding of the models' dynamics. Static reservoirs are in a somewhat opposite situation: by combining fixed weights with linear readouts, their dynamics and tuning mechanisms are better understood and easier to implement (Jaeger, 2002; Lukoševičius and Jaeger, 2009; Lukoševičius, 2012), but their simplicity suggests that further architecture improvements, especially regarding the random initialization of weights, are possible and desirable.

As we have seen in Chapter 4, improvements in multiple information processing capacities are associated with dynamics near criticality in multiple

---

[1] The development of reservoir computing, in fact, was aided by the backpropagation-decorrelation algorithm (Steil, 2004; Lukoševičius and Jaeger, 2009), which decouples adaptive rules for recurrent and output layers, i.e. it trains a separate readout layer as in reservoir computing frameworks.

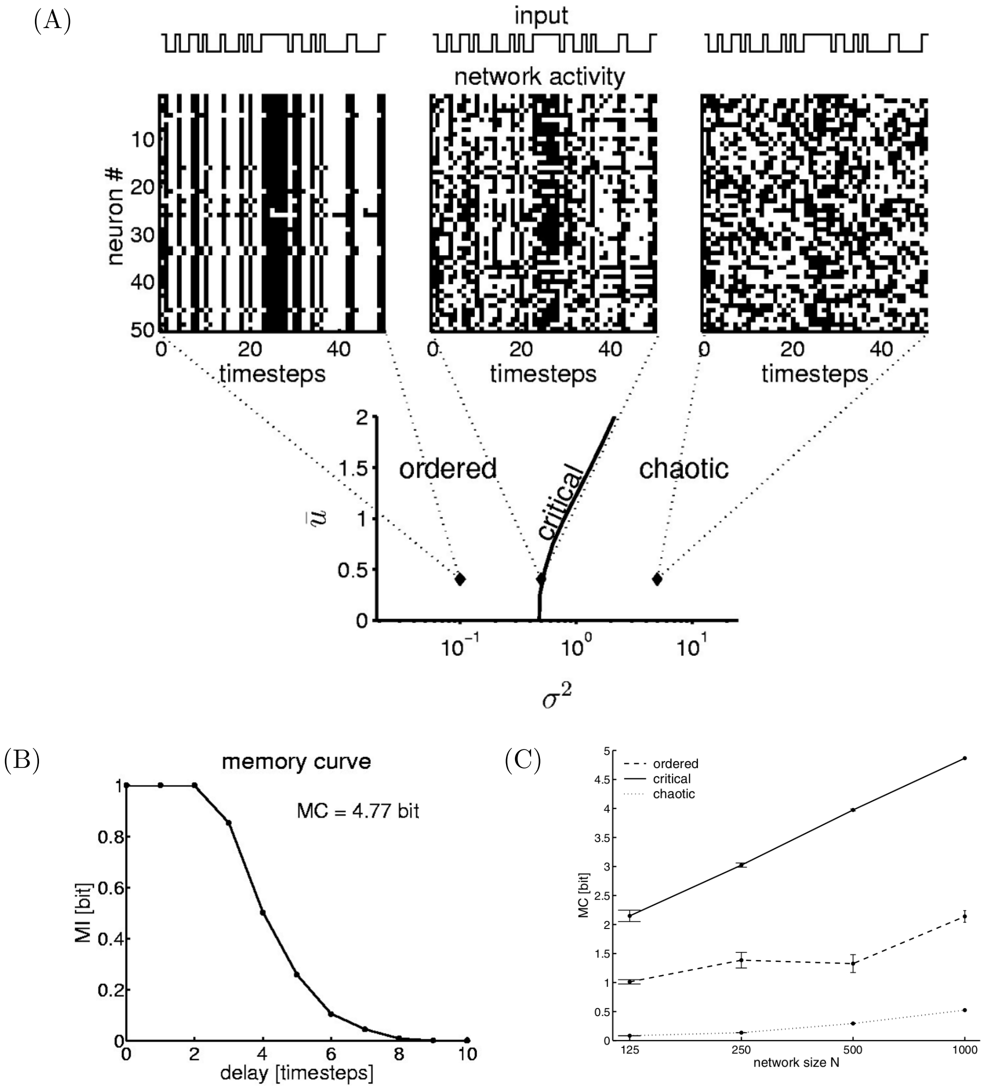


Figure 5.1: **Fading memory scaling for distinct dynamics in reservoirs (Bertschinger and Natschläger, 2004).** (A) Ordered, critical, and chaotic dynamics in externally driven recurrent networks. The dynamical regimes were tuned via an external input (u) and weight distribution (Gaussian with zero mean and σ^2 variance). (B) Memory curve for a three-bit parity task. The mutual information between input and output, MI, decayed for increasing time delays between input and prediction. The memory capacity, MC, was estimated as the total information retained by the network (area under the curve). (C) Scaling of memory capacity. In contrast to other states, critical dynamics yielded logarithmic growth (curves show averages of 10 randomly drawn networks for each state). Figures were reproduced with permission from the original manuscript.

neural network models and experimental setups (Shew and Plenz, 2013). It is, therefore, reasonable to hypothesize that the same occurs for static reservoirs and randomly connected recurrent neural networks. In particular, in networks with fixed connections, the dynamical state is dictated by their connectivity pattern, weight distributions, and external drive, all of which can be chosen in order to tune the model to a phase transition point. Interestingly, this hypothesis has been verified for a particular type of reservoirs: externally driven networks of gated units and fixed, randomly chosen connections[2] can be poised at a phase transition point with critical dynamics and optimized computation abilities (Bertschinger and Natschläger (2004); Fig. 5.1A). Importantly, in this study, criticality, or edge-of-chaos dynamics, was defined in terms of small perturbations and Lyapunov exponents, via the estimation of the Hamming distance for infinite systems (or equivalently, a mean-field approximation for the thermodynamic limit). As we discussed in the last chapter, edge-of-chaos dynamics is, in principle, a distinct phenomenon and does not necessarily co-occur with power-law distributed neuronal avalanches (Kanders et al., 2017a), and thus we differentiate between *edge-of-chaos criticality* and *avalanche criticality*. Although the authors of Bertschinger and Natschläger (2004) did not investigate the occurrence of avalanche criticality, they have shown that edge-of-chaos criticality supports complex computations in randomly connected neural networks, by examining a three-bit parity task[3]. The performance of the network was measured as the mutual information between the true input parity and the network output (given by a linear readout), which decayed as a function of time delays between input and prediction (Fig. 5.1B). From this decay, or memory curve, an overall fading memory capacity could be estimated as the amount of information on past inputs a network was capable of retaining. This memory capacity was shown not only to be maximal for networks at edge-of-chaos criticality but also to scale logarithmically with the network size, while networks with ordered and chaotic dynamics showed a much less steep growth (Fig. 5.1C). The result demonstrated that reservoirs indeed

---

[2] Note that such networks are qualitatively similar to a SORN composed of only excitatory neurons, without plasticity mechanisms or membrane noise. The main difference is their spread of a single input among all units, while SORNs restrain distinct inputs to distinct subgroups of neurons.

[3] A parity task consists in a task whose desired output network signal should be given by the parity of three past binary inputs, $u(t - t_p)$, $u(t - t_p - 1)$, and $u(t - t_p - 2)$, for increasing delays t_p.

have increased information processing capacities when tuned to edge-of-chaos criticality (although not necessarily avalanche criticality). Thus, adaptation mechanisms that are able to always drive reservoirs towards this state can be particularly useful for their memory properties. Given the appearance of avalanche criticality in SORNs, we again turn our attention to their plasticity mechanisms and investigate possible links between self-organization towards criticality and improved memory capacity.

5.1.1 Self-organization in reservoirs

In order to improve a static reservoir's performance and memory, different tuning techniques have been proposed, with various degrees of success. For example, multilayer readouts (Maass et al., 2002), multiple readout supervised training techniques (logistic regression, pseudo-inverse methods, ridge regularization; Lukoševičius and Jaeger (2009)), readouts trained via reinforcement learning (Legenstein et al., 2008), evolutionary algorithms (Xu et al., 2005), clustering (Bush and Anderson, 2006), and support vector machines (SVMs; Shi and Han (2007)) exist in the literature. Instead of reviewing and comparing each of these readout training methods, we here focus on the update of a reservoir's weights, with the goal of studying the effects of self-organization towards an (edge-of-chaos) critical dynamics on its learning abilities. We take advantage of the biologically inspired plasticity mechanisms from the SORN model, which were shown to originate and maintain (avalanche) criticality signatures (Chapter 4). Thus, our goal is not only to show how plasticity driven dynamics can generally improve reservoirs but also link the appearance of edge-of-chaos criticality in self-organized reservoirs with a potential function of criticality in the brain: the improvement of the fading memory capacity.

The original $\mathrm{SORN_L}$ model already established that a combination of spike-timing-dependent plasticity (STDP), synaptic normalization (SN), and intrinsic plasticity (IP) outperforms static reservoirs in various pattern learning tasks (Lazar et al., 2009)[4]. Interestingly, however, the model displayed subcritical dynamics under small perturbations (Lyapunov exponents estimated via Hamming distance — thus, the edge-of-chaos definition of criticality) when subject to structured, repeating external input of learning tasks. In order to better understand how this combination of plasticity mechanisms

---

[4] For a detailed description of SORN models and their variations, see Chapter 3.

can improve a model's memory, potentially due to self-organization towards a critical state, we compared the fading memory capacity MC of SORNs and static reservoirs driven only by random external input, in the form of random sequences of symbols (details on the fading memory estimation are provided in the next section). Importantly, the lack of temporal structure in the external input avoided any possible interference of spatio-temporal learning in the model's dynamical regime and/or fading memory estimation. As expected, synaptic plasticity improved the overall memory capacity after a short self-organization period (approximately $T_{\text{plast}} = 10\,000$ time steps), which was not observed in random networks (i.e., static reservoirs with excitatory and inhibitory units) or in SORNs with only STDP or IP (Fig. 5.2). Additionally, for a network of size $N^{\text{E}} = 100$, the memory capacity after self-organization reached the same level as a reservoir twice its size, confirming that synaptic plasticity is indeed a powerful mechanism to improve a reservoir's memory capacity.

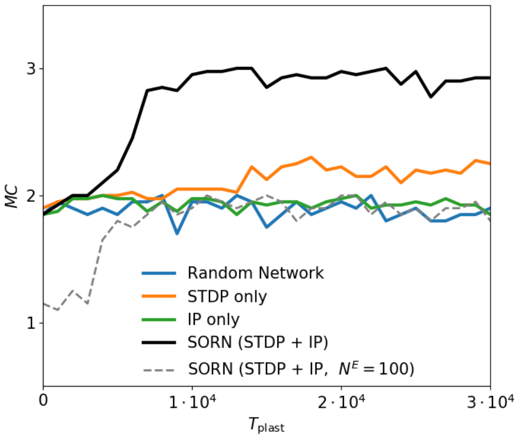


Figure 5.2: **Self-organization due to plasticity improves the fading memory capacity.** Fading memory capacity (MC) as a function of convergence time (T_{plast}) for static reservoirs (Random Network) and SORNs with and without STDP and IP plasticity rules ($N^{\text{E}} = 200$). The dashed gray line shows the convergence for a smaller SORN, whose memory capacity is comparable to a static reservoir twice its size.

5.2 Fading memory capacity

The fading memory capacity, which was independently named *short-term memory* in echo state networks (Jaeger, 2002), is the ability of a system to retain and combine information about previous inputs in its current activity, at any given time step. In the case of recurrent neural networks, this capacity can be estimated from the state $\mathbf{x}(t)$, when such state is a function of a finite number of past inputs $u(t - t_p)$, for t_p in some time window $[-T, 0]$ (Maass et al., 2002)[5]. Such property allows for appropriate readout functions to decode past recent inputs from the network state[6], what underlies the complex temporal pattern learning abilities of recurrent neural networks. Importantly, the fading memory capacity is essentially different from a network's learning capacity: while the first stores information about the input *history* in the recurrent activity, the second refers to the encoding of information about a specific task and its parameters. In fact, recent work has shown that various recurrent neural network architectures (such as LSTMs and GRUs) achieve nearly the same general capacity bounds (per task and per unit), which varied only due to distinct training methods' effectiveness (Collins et al., 2016). Although task information encoding increased approximately linearly with the number of parameters of the tested models, the authors showed that input history information increased instead with the number of hidden units, suggesting that the fading memory capacity might not depend on a particular task learning capacity. The extent to which these separate properties relate in different models or depend on distinct dynamical states, however, is currently unknown.

In order to estimate the fading memory capacity of static reservoirs and SORNs, we took advantage of a modified Sequence Learning task, which was already employed in the last chapter as an example of structured external input. The input of the Sequence Learning task consisted of a repeating sequence, with length L, of symbols drawn from an alphabet of size $\mathrm{U_A}$. Each symbol provided extra input to a fixed, randomly chosen subset $N^{\mathrm{U}} < N^{\mathrm{E}}$ of reservoir units or excitatory neurons. The readout layer was trained on the network activity vector $\mathbf{x}(t)$ for T_{train} time steps while all plasticity mechanisms were turned off. Subsequently, the performance of the model was evaluated for T_{test} time steps, again with plasticity mechanisms off, but following distinct procedures for the estimation of learning or fading memory capacities. For learning tasks, as described previously, the performance was defined as the percent of correct predictions of the *next* input symbol. For

---

[5] This property implies that the initial state is forgotten in a finite number of time steps, which makes the fading memory capacity, by definition, finite (Lukoševičius and Jaeger, 2009).

[6] Note that, for a model without fading memory capacity, this must not necessarily happen. The network state could retain information about initial conditions or noise (in case of a non-deterministic system). In these cases, training a readout layer to recover information about recent inputs is more difficult or impossible. As we will see later in this chapter, this is exactly what happens for a supercritical, noise-driven dynamics.

memory tasks, however, the input consisted of non-repeating random sequences ($L \to \infty$) and the performance was evaluated on past time steps: at each time step t, the performance was defined as the normalized number of correct classifications of input symbols received t_p time steps in the past. Again, we emphasize that both capacities refer to different phenomena: the learning capacity measures the encoding of temporal patterns from the input sequence, while the fading memory capacity measures the amount of past information a model can retain in its recurrent activity. This justifies our choice for random sequences as input for memory tasks, as they contain no temporal pattern to be encoded, thus allowing for fading memory capacity estimation.

5.2.1 Critical capacity scaling with plasticity action

We defined the fading memory capacity MC as the average memory of recent symbols stored in the $\mathrm{SORN_L}$'s excitatory activity after self-organization, i.e., the average number of past time steps in which past input symbols could be correctly classified by the readout layer with a reasonably small error. In order to define what a reasonably small error was, we first looked at the memory curves for networks of various sizes (see Fig. 5.3A and Fig. 5.3B for examples of memory curves for $N^{\mathrm{E}} = 100$ and $N^{\mathrm{E}} = 1600$, respectively). The curves showed that the error quickly increased as a function of the delay t_p for different alphabet and network sizes, and eventually reached chance level. As expected, bigger networks were able to recall inputs for longer delays, and bigger alphabets[7] resulted in seemingly smaller memory capacities due to faster memory decay. Interestingly, networks of size $N^{\mathrm{E}} = 1600$ could recall input symbols up to 10 past time steps with a performance of around 90%, when the alphabet was small enough ($\mathrm{U_A} = 10$).

Given the quick increase in error (or decay of the memory), a natural definition of reasonably small error was a fixed threshold that could detect how many past time steps would cause such decay. Based on the memory curves of a network of size $N^{\mathrm{E}} = 200$ (Fig. 5.3C), we set this threshold at 10% (gray dashed line). Thus, we estimated the fading memory capacity as

---

[7] Note that for big alphabet sizes and small networks, the input neuron pool for different symbols necessarily had overlaps, i.e., there existed neurons that received input from two or more distinct symbols. Our tests (not shown) suggested that these overlaps did not have significant effects on the model's performance as long as they were kept to a minimum and any two symbols never had identical input pools.

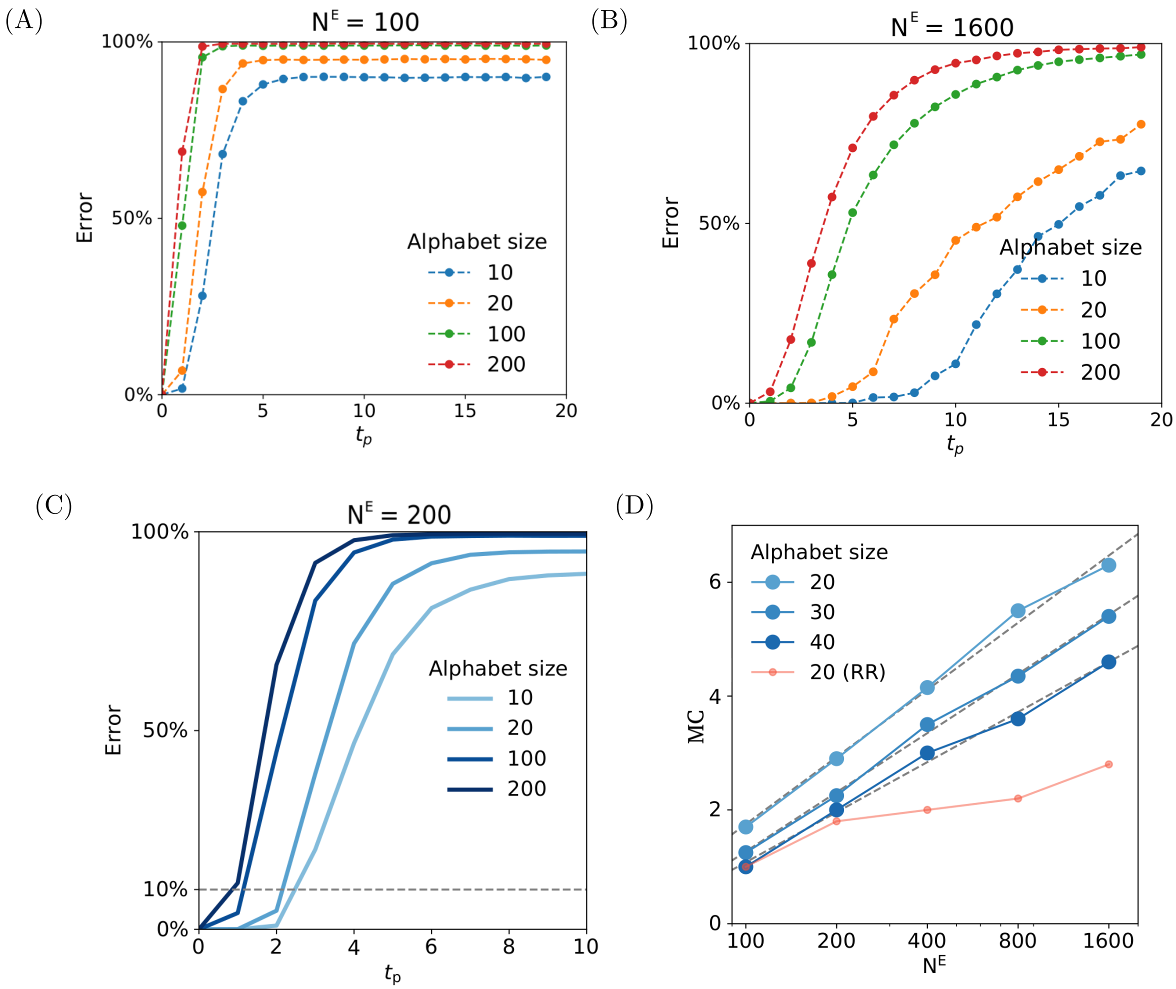


Figure 5.3: **Memory curves and memory capacity scaling for the $\mathbf{SORN_L}$.** (A) Memory curves for a network of size $N^{\mathrm{E}} = 100$, for different alphabet sizes. Curves show the average classification error increase for increasingly long delays, until stabilization at around chance level. (B) Equivalent curves for $N^{\mathrm{E}} = 1600$. (C) Memory capacity definition via error threshold (gray dashed line) for $N^{\mathrm{E}} = 200$. (D) Fading memory MC scaling with network size, for different alphabet sizes. Gray dashed lines show logarithmic functions fitted to the data (of the form $a \cdot log(N^E) + b$; $A = 20$: $(a, b) \approx (1.70, -6.09)$; $A = 30$: $(a, b) \approx (1.50, -5.64)$; $A = 40$: $(a, b) \approx (1.27, -4.78)$). Red curve shows the scaling for a random static reservoir (RR), which has a slower increase compared to logarithmic functions. All plots show averages of 10 independent simulations (standard deviations have been omitted for the sake of clarity).

the average number of past time steps from which a network could recall inputs with performance higher than 90%. Note that, since the time steps were discrete, we typically interpolated linear curves between consecutive points in order to achieve a more precise estimation. Remarkably, the fading memory capacity increased approximately logarithmically with the network size (Fig. 5.3D), independently of the alphabet size, reaching $MC \approx 6$ for $N^{\mathrm{E}} = 1600$ and $\mathrm{U}_{\mathrm{A}} = 20$[8]. Such scaling, which was not seen in static random reservoirs, was comparable to the one observed in recurrent networks operating at the edge-of-chaos (compare Fig. 5.3D to Fig. 5.1C), suggesting that self-organization due to plasticity action might be as beneficial to reservoirs as critical dynamics.

5.2.2 A link between neuronal avalanches and memory improvement

We have seen that self-organization due to plasticity mechanisms in the SORN is capable of originating avalanche criticality signatures in its spontaneous activity, but that they disappear under structured input (Chapter 4). Now, we have additionally shown that plasticity can also improve the fading memory capacity of reservoirs and yield a logarithmic scaling only observed during edge-of-chaos criticality. It remains to been verified, however, whether a fading memory improvement is also observed at the dynamical state in which power-law distributed neuronal avalanches appear. In principle, states in which neuronal avalanches are present do not need to show optimal memory capacity. Networks of spiking neurons with fixed synapses, for example, show a peak in information storage at a state with a different configuration than the one where neuronal avalanches are observed (Mediano and Shanahan, 2017). Interestingly, nonetheless, the latter state balanced information storage and transfer, suggesting that power-law distributions can be observed in a state that, although not optimal for a single information processing capacity, balances the performance of important brain functions. Additionally, recent studies have suggested that near-critical dynamics are necessary to stabilize and consolidate sparsely driven input representations (Skilling et al., 2017), offering further support to a functional role of criticality in the

---

[8] Note that a fading memory capacity of 6 for an alphabet size of 20 means that the network needs to be able to distinguish $20^6 = 64$ Million patterns with an error smaller than 10%!

memory capacity of a network.

Our results on the SORN model go in the same direction, suggesting a link between improved memory and neuronal avalanches. We tested whether the power-law distributed neuronal avalanches present in the SORN_Z's spontaneous activity were associated with an improvement in its fading memory by comparing different levels of membrane noise during the Sequence Learning task with random sequences as input ($L \to \infty$). Three different noise levels, which had revealed the existence of an externally driven phase transition point (see section 4.2.4), were compared[9]. For consistency, we stick here to the same nomenclature as the previous chapter: low, medium, and high noise levels correspond to Gaussian membrane noise with zero mean and variance $\sigma^2_\text{L} = 0.005$, $\sigma^2_\text{M} = 0.05$, or $\sigma^2_\text{H} = 5$, respectively. High levels of membrane noise abolished the network's learning and memory capacities, and the performance remained just slightly above chance level (Fig. 5.4). Such low performance was expected, as any input information quickly gets lost in a purely stochastic state. Medium noise levels, despite their higher stochasticity, were not only associated with neuronal avalanches but also resulted in improved fading memory capacity, as inferred from its improved memory curve, in comparison to low noise levels. This rather counter intuitive result suggested that criticality signatures in spontaneous activity might indeed be associated with an increase in the computational power of networks for tasks requiring the recognition of temporal patterns.

Taken together, our simulations showed that both "types" of criticality do not co-occur in SORN models after self-organization. Interestingly, however, even though criticality signatures are not associated with maximum performance in relatively simple learning tasks, we highlight that they coincide with an improvement of the time scale of memory persistence in the models. Furthermore, such capacity scaled logarithmically with the network size, a property so far observed only in reservoirs tuned to edge-of-chaos criticality (Bertschinger and Natschläger, 2004). These findings imply that the presence of criticality signatures might not only be beneficial to a network's memory but also linked to the existence of edge-of-chaos critical dynamics, as both phenomena result from an increase in the long-range correlations across units and in time. Importantly, this implies that the dynamical state of the cortex or of a neural network could (and ideally should) be adapted to task

---

[9] The membrane noise was provided to the neurons in addition to the input from the Sequence Learning task, resulting in a combined non-deterministic neuronal input.

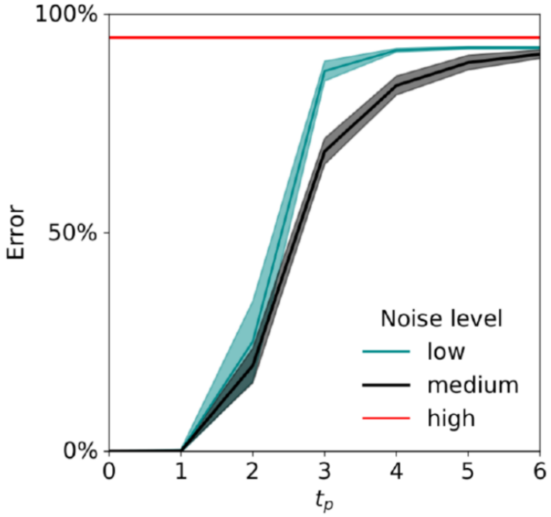


Figure 5.4: **Memory curves show an improvement in the memory capacity at medium noise levels.** Error in input recall as a function of the delay time t_p for a SORN$_Z$ of size $N^E = 200$, when performing a Sequence Learning task (alphabet size $A = 20$). Smallest errors were observed for medium noise levels, when neuronal avalanches occur. Curves show averages of 20 simulations, and shaded areas represent the standard deviation.

requirements. Such tuning is particularly sensitive if the network operates in a reverberating state in the vicinity of criticality, as similar network models have recently shown (Wilting et al., 2018).

A formal description of the relation between those tuning properties, however, is beyond the scope of this thesis. Such a description necessarily requires a better understanding of the classes of systems in which edge-of-chaos and avalanche criticality co-occur, or at least are correlated to some extent. The investigation of these classes is still in its infancy, and future research on criticality should shed light on the relationship between distinct "critical" points and their relation to functional properties in networks. Instead, we focus now on our final topic of investigation in this work: a practical application of self-organizational improvements to learning and memory capacities of recurrent neural networks, in the form of simple grammar learning and language processing.

5.3 Simple grammar learning: "novel" sentence generation

After learning repeating sequences of symbols with a fixed length, a reasonable next step in terms of task complexity is learning sequences that combine various temporal dependencies. Such sequences, for example, may integrate long and short-range temporal dependencies, in which each symbol can only be correctly determined by recalling information from multiple, not necessarily consecutive, past time steps. A typical example is language: in order to be correctly placed in a given sentence, a character (or even a phoneme) depends on the near past (for example, the current word), the previous words in the sentence, and finally on the context in which the sentence is inserted. Combining all these time scales, although challenging, is a task which biological neural circuits naturally excel. However, the mechanisms underlying these processes are still not fully understood. Taking advantage of the SORN sequence learning framework and self-organization mechanisms, we assessed whether these models can learn simple grammar rules by training the network on sentences with grammatical structures and investigating its posterior self-generated sentences.

Based on its performance on sequence learning, we chose the SORN_L for the grammar learning tasks. As previously, the model was trained with sequences of sentences, one character per time step, during T_plast time steps. Each sentence, built by selecting at random (with equal probability) words from a predefined dictionary and combining them according to a fixed grammar, contained a full stop character (end-of-sentence marker) and started with a space character (both considered independent symbols, which were treated as normal letters), while all other characters were kept lowercase for simplicity. Subsequently, synaptic plasticity was frozen and the readout layer was trained for T_train steps in order to predict the *next* character in the input (as previously, the readout training was performed via logistic regression). The addition now was the autonomous phase: after the readout training, external input was replaced by the output of the readout layer (i.e., the next character prediction) for T_auto time steps, creating a feedback loop, while plasticity was still kept frozen (Fig. 5.5). This turned the network into an autonomous dynamical system that generated sequences of letters and sentences, which could then be classified as correct or incorrect according to some definition of error type. The input sentences were drawn from predefined dictionaries (see Appendix C for more details and an overview of all the sentences contained in each dictionary) and the total training time steps were set to $T_\text{plast} = 100\,000$ and $T_\text{train} = T_\text{auto} = 10\,000$.

We first tested whether the model was able to learn and generalize grammar rules from a subset of dictionary sentences with the same structure. The sentences consisted of tuples of words of various lengths, according to the pattern "[*subject*] [*verb*] [*object*]." (including spaces and the full stop).

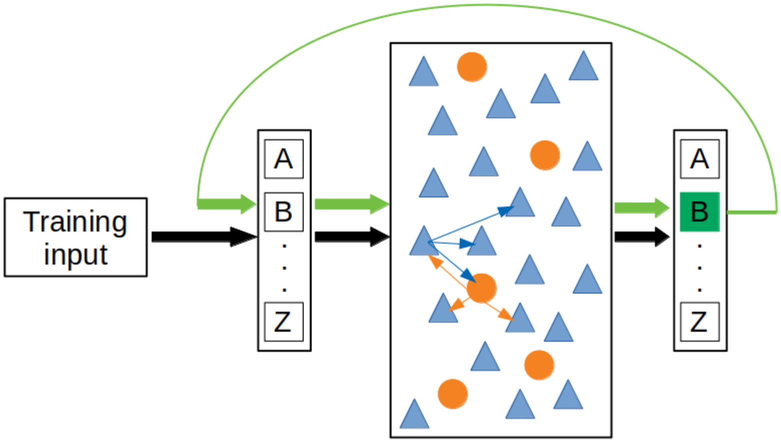
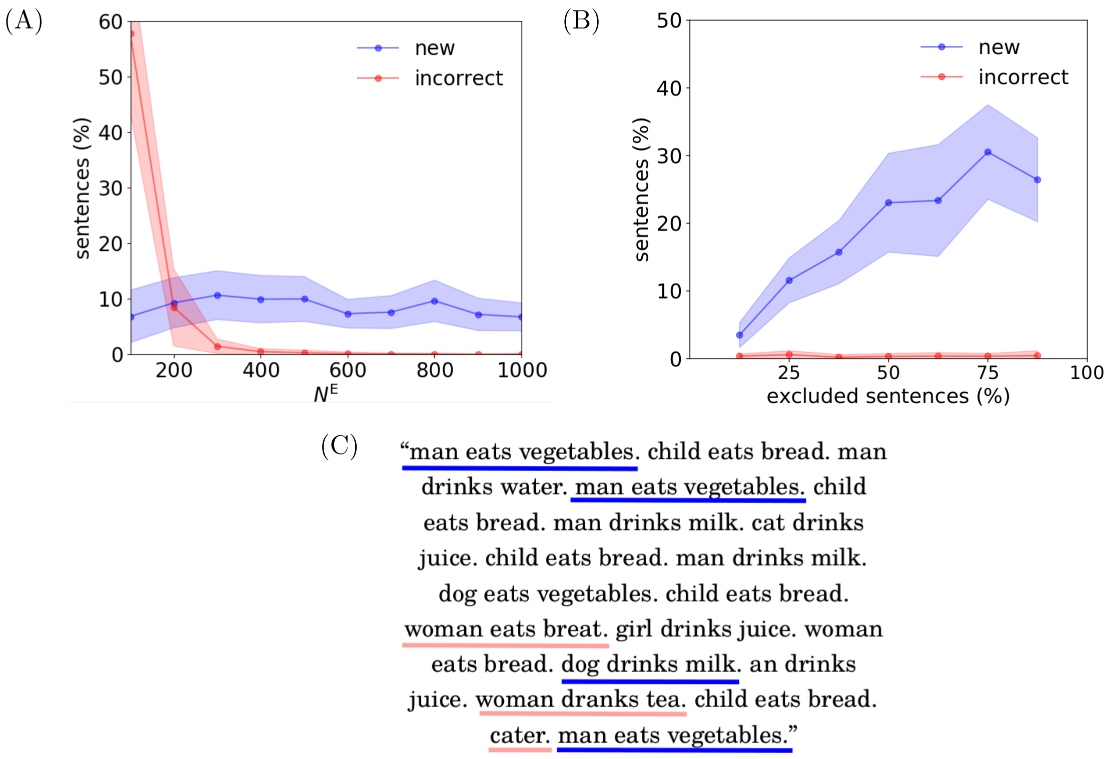


Figure 5.5: **Grammar learning task: training and autonomous phases.** During the training phase, the SORN_L receives training input, with which the reservoir and the readout layer are trained separately (for T_plast and T_train time steps, respectively; black arrows). Afterwards, the training input is cut off and the output (next character prediction) becomes the new input, creating a feedback loop (green arrows), while plasticity is kept frozen.

Combinations of subjects, verbs, and objects were randomly drawn from a dictionary and were constrained by a simple rule: objects could only be correctly associated with one of two verbs. The predefined dictionary for this task, FDT[10], contained 8 subjects, 2 verbs, and 8 objects (4 associated with each verb), and thus was able to generate a total of 64 distinct sentences, with an alphabet size (including letters, spaces, and punctuation) of $\text{U}_\text{A} = 25$. During the network training phase (including the readout training), a fixed number of randomly chosen sentences were excluded from the input in such a way that every word would, necessarily, still appear at least once. The performance was estimated by a comparison of three classes of output sentences (i.e., all the characters separated by full stops) in the autonomous phase: *correct*, sentences that had appeared in the input; *incorrect*, sentences that could not be generated by the FDT dictionary and contained incorrect combinations of words or characters; and *new*, sentences that could be constructed from FDT, but were excluded from the input. We quantified those types of

---

[10] This dictionary was named after its most iconic sentence, “ fox drinks tea.”.



"man eats vegetables. child eats bread. man drinks water. man eats vegetables. child eats bread. man drinks milk. cat drinks juice. child eats bread. man drinks milk. dog eats vegetables. child eats bread. woman eats breat. girl drinks juice. woman eats bread. dog drinks milk. an drinks juice. woman dranks tea. child eats bread. cater. man eats vegetables."

Figure 5.6: **Generation of new sentences by the $\mathbf{SORN_L}$** (A) Incorrect and new sentences as a function of the network size for the FDT dictionary, when 25% of the possible input sentences were excluded from training. Curves show averages of 20 independent simulations and the shaded area shows the 25% to 75% percentile interval (with linear interpolation between points). Although the percent of new sentences fluctuated around 10%, incorrect sentences were virtually absent for networks containing more than 600 excitatory neurons. (B) Incorrect and new sentences as a function of the relative number of excluded input sentences, for the same dictionary and $N^{\mathrm{E}} = 400$. New sentences peak at 75% due to the FDT definition (see text). Shaded areas show the same interval as (A). (C) Sample autonomous output sentences for $N^{\mathrm{E}} = 200$. Blue and red lines highlight new sentences, which did not appear in the input, and typical incorrect sentences, respectively.

output by estimating the percent of the *total* number[11] of output sentences that could be classified in each class[12]. Interestingly, the relative number of incorrect sentences quickly decreased with network size N^{E} and was overtaken by the relative number of new sentences in networks composed of more than approximately 200 excitatory neurons, when 1/4 of the dictionary sentences were removed from the training input (Fig. 5.6A). The relative number of new sentences roughly fluctuated around 10%, independently of the network size, while the relative number of incorrect sentences was virtually zero for $N^{\mathrm{E}} > 600$. This result was surprising since, after random initialization, the $\mathrm{SORN_L}$ turned into a deterministic model, and thus the new sentences were generated purely as a result of its unsupervised input encoding[13]. This showed that the model was indeed capable of learning distinct temporal patterns, outputting correct characters and combinations of words of various lengths. The relative number of new sentences, furthermore, was limited not by the model's learning capacity, but instead by the number of excluded input sentences. An increase in the percent of excluded input sentences led to a roughly linear increase in the relative number of new sentences up to a maximum point, while the number of incorrect sentences remained negligible (Fig. 5.6B). As expected, the maximum peak for new sentences, at around 75%, was a consequence of the FDT dictionary definition: when more than 75% of the sentences were excluded, the input did not contain enough variability to allow for generalization, and new sentences were not as frequently observed in the output.

A detailed look at the incorrect sentences provided additional insights into the temporal dependencies the model was able to learn (Fig. 5.6C). First, spaces and full stops typically appeared at the expected correct places, including for new sentences. Second, most errors consisted of single incorrect characters (as in "breat" instead of "bread" or "dranks" instead of "drinks"). Third, a few incorrect sentences seemingly consisted of a mixture of distinct dictionary words sharing a single letter, as in "cater" (possibly a combination of "cat" and "water", both contained in the FDT). These types or errors suggested that the model, due to self-organization, was indeed able to encode some letter statistics from the input, as the output contained letters that were

---

[11] In practice, we removed the first and last output sentences from the analysis, as they could be incomplete, depending on the number of time steps in the autonomous phase.

[12] Note that the classes are exhaustive and mutually exclusive by construction, thus their percentages add up to 100%.

[13] Recall that during the autonomous phase all plasticity mechanisms were kept frozen.

generated, sometimes incorrectly, based on the observed temporal patterns.

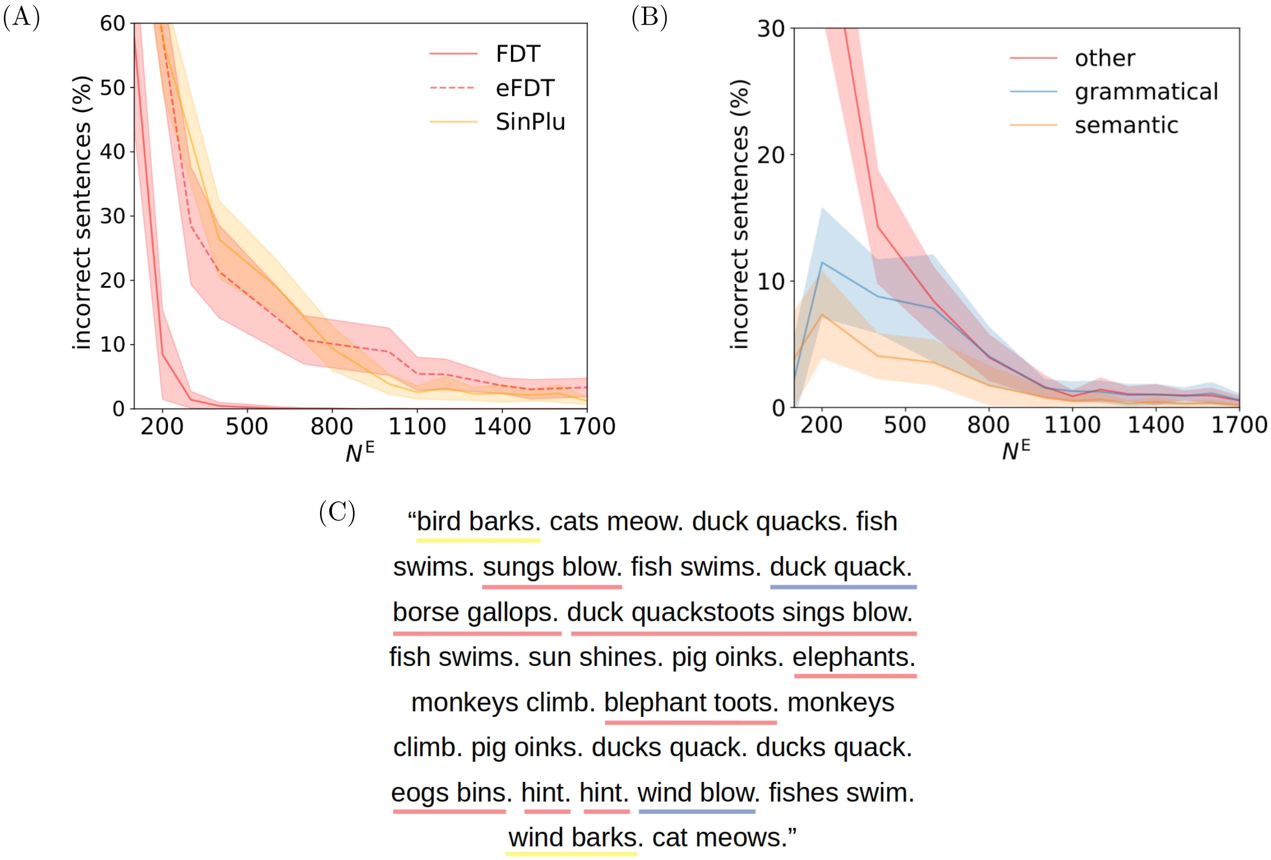
The SORN_L's capacity of encoding input statistics at the character level suggested that more complex tasks could potentially be performed. Given the nature of plural construction rules in the English language, in which most plural nouns and verbs consist of their singular forms with the addition or removal of the letter "s", we extended the FDT dictionary to include singular and plural subjects and verbs, maintaining the constraint for verb and object association. Learning and correctly differentiating singular and plural sentences could be particularly challenging if the model relied only on short term temporal dependencies, since the correct placement of the characters depends on previous words[14]. The extended FDT, or eFDT, was composed of 128 distinct sentences, including their singular and plural forms (see again Appendix C for details). Additionally, we considered yet another simplified dictionary containing singular and plural "[*subject*] [*verb*]." pairs (e.g., "dog barks." and "dogs bark.") for 12 sentences, SinPlu (Appendix C). SinPlu was relevant not only due to the "s" placement, which necessarily relied on longer temporal dependencies, but also due to its similarity to morphological parsing[15], which has been typically employed in statistical learning models of natural language and speech transcription (Willits et al., 2009; Huebner and Willits, 2018).

Similarly to the FDT, the autonomous output of the SORN_L after training on the eFDT and the SinPlu dictionaries was evaluated by looking at the relative percentages of incorrect sentences. Here, however, we did not remove any sentences from the input, and thus the model could not generate novel correct ones. Not surprisingly, both dictionaries resulted in a bigger number of incorrect sentences when compared to the FDT output (Fig. 5.7A), confirming that tasks including singular and plural sentences are generally more complex due to the higher ambiguity of input letters. Interestingly, the relative number of incorrect sentences decayed very similarly for the eFDT and the SinPlu, reaching a minimum of around 3% for $N^\text{E} > 1500$. For the SinPlu output, we further differentiate incorrect sentences as *grammatical*, in which

---

[14] For instance, in the sentence pair "cat meows." and "cats meow", the correct placement of the letter "s" after the verb depends on the previous occurrence of the same letter after the subject.

[15] Morphological parsing refers to the splitting of word ending tokens, such as the plural forms "s" and "es" or the past-tense "ed", into new words in order to facilitate learning at the word-level. This is equivalent to the split of one character, "s", from the rest of the word, which happens by construction in character-level leaning models.



"bird barks. cats meow. duck quacks. fish swims. sungs blow. fish swims. duck quack. borse gallops. duck quackstoots sings blow. fish swims. sun shines. pig oinks. elephants. monkeys climb. blephant toots. monkeys climb. pig oinks. ducks quack. ducks quack. eogs bins. hint. hint. wind blow. fishes swim. wind barks. cat meows."

Figure 5.7: **Errors in the generation of singular and plural sentences.** (A) Percent of incorrect output sentences as a function of the network size, for different artificial dictionaries. In contrast to FDT, eFDT and SinPlu contained singular and plural pairs of sentences, which are harder for a model to learn (see text). (B) Types of incorrect sentences for the SinPlu dictionary. Grammatically incorrect sentences are more common than semantically incorrect ones for small networks, but their numbers greatly decrease with network size. Curves show averages of 20 independent simulations, and the shared area show the 25% to 75% percentile interval (with linear interpolation between points). (C) Sample autonomous output sentences after training on SinPlu, for $N^{\mathrm{E}} = 200$. Colors indicate the types of errors: grammatical (blue), semantic (yellow), and other (red).

the subject and the verb were correct but their number did not agree due to the position of the letter “s” (e.g. “dog bark” or “dogs barks”); *semantic*, in which the subject and the verb agreed in number but their combination was not present in the input dictionary (e.g. “dog meows”); and *other*, for the remaining incorrect sentences (e.g. “dog arks”; Fig. 5.7B). As additional evidence of the higher complexity of the task of differentiating singular and plural forms, the grammatically incorrect sentences were systematically more numerous than the semantically incorrect ones, especially for small networks, although both were outnumbered by the other types of errors (Fig. 5.7C.

In conclusion, the SORN’s ability to generate new, correct sentences according to predefined grammar rules was remarkable since its dynamics were deterministic and its reservoir learning rules, STDP, IP, and SN, were unsupervised. Importantly, the capacity of encoding various time dependencies in its activity, which was evidenced by the learning of singular and plural forms of sentences, was a potential consequence of the improved $\mathrm{SORN_L}$’s fading memory capacity, which itself was a result of the biologically inspired self-organizational mechanisms that we extensively discussed in the previous chapters and sections. Thus, our results on simple grammar learning suggested that self-organization in neural networks can result in powerful learning abilities even on relatively small models with a few hundred or thousand neurons. Such learning abilities were particularly encouraging for both reservoir computing and plasticity driven self-organization and might pave the way for future detailed optimization of unsupervised information encoding via biologically inspired plasticity mechanisms. For now, in the last part of this study, we briefly investigated a more realistic task, focusing on natural language acquisition.

5.4 SORNs as generative language models

Temporal patterns from artificially constructed dictionaries, of course, are distant from the complexity and noise of natural language. Furthermore, language learning has been suggested to rely not only on temporal patterns but also on acquired semantic knowledge (Huebner and Willits, 2018), a feature that is absent from most recurrent neural network models. Although the mechanisms responsible for the emergence of semantic knowledge are difficult to identify and currently not fully understood, infants seem to use computational strategies to detect statistical patterns in language input, which are learned and reproduced with remarkable speed (Kuhl, 2004). Even more surprisingly, this learning process occurs simply by exposure to speech, oftentimes without supervision, a process that is potentially analogous to the unsupervised reservoir training methods we have been discussing so far. Given the promising results on simple grammar rules, we next investigate how SORN models perform on language learning tasks, and discuss whether a deterministic neural network model with biologically plausible learning rules can perform statistical learning and generate sentences that resemble a highly variable realistic input.

5.4.1 Language acquisition

Even though the underlying neural mechanisms of language acquisition are not yet completely understood, their study has been of great interest, in particular, because infants are able to learn to speak even without supervision, just by being exposed to speech. It has not only been shown that infants learn language through some form of statistical inference (Kuhl, 2004), but also that the language outcome of children is highly correlated with the amount and variety of language input they received as infants (Golinkoff et al., 2015). Furthermore, 8-months old infants are already able to extract words from exposure to fluent speech from an artificial language, where the only cues are the conditional probabilities between syllables within words compared to syllables between word boundaries (Saffran et al., 1996). Such excellent language learning skills are, as other general learning abilities, determined by multiple plasticity mechanisms in the brain, thus the use of computational models could be essential for the clarification of their roles and capabilities.

In order to study the $\mathrm{SORN_L}$'s behavior under the input of realistic language, we employed the American English CHILDES database (MacWhinney, 2000), as precompiled by Willits and colleagues (Willits et al., 2016). The dataset contained transcripts of interactions between parents and infants aged up to 72 months in various situations (see Appendix C for more details and an overview of the required pre-processing methods). With simpler vocabulary and shorter, repetitive words, this dataset contained features and short temporal patterns which we expected the SORN to capture, in an analogy to language acquisition in infants. We employed a training procedure identical to the previous grammar learning tasks, training the network with unsupervised synaptic and homeostatic plasticity mechanisms and a separate supervised readout layer to predict the next letter in the input corpus (for $T_{\text{plast}} = 500\,000$ and $T_{\text{train}} = 30\,000$, respectively). After training, and with all plasticity frozen, this prediction was again fed back as input, resulting in an autonomous system that generated potentially "new" speech-like words and sentences. Again, as our model was built at the character-level, we treated spaces and punctuation as single characters, which resulted in an alphabet size of $\mathrm{U_A} = 32$ (rare, non-ASCII characters were removed, while all letters were converted to lower case).

Importantly, the performance of a language generation task based on transcripts of speech is not trivially measured. Much like real language, and differently from our artificially generated grammars, a full set of correct possible sentences does not exist, and even the input contains various patterns and words that could be considered correct or incorrect under different assumptions[16]. Thus, instead of searching for new grammatically correct sentences in the spontaneous output, we first described the performance of the $\mathrm{SORN_L}$ model in predicting the *next* letter from the input (Fig. 5.8A), by comparing models of different sizes while keeping the input exactly the same. Specifically, the plasticity phase always received the first 500 000 characters from the CHILDES input corpus, the readout layer was trained on the subsequent 30 000 characters, and finally evaluated on the next 30 000. As expected, the average prediction performance for each letter increased with network size for the majority of characters, although even relatively big networks (with tens of thousand neurons) were not capable of reaching a high performance for the majority of them. One notable exception was the space character, which separated words. Its performance was comparatively higher as the network only had to identify the time step in which a word ends, which did not require a long memory scale. In particular, output samples showed that smaller networks tend to generate shorter words on average, which resulted in a higher count of, and consequently a higher performance for space characters.

The fluctuating performance observed for the remaining letters could be potentially a consequence of their frequency in the input dataset. We compared the frequency of characters from CHILDES to the ones typically observed in the spontaneous output of the $\mathrm{SORN_L}$ (Fig. 5.8B). This comparison revealed an important difference: although the distributions generally followed a similar decay (i.e., most frequent input characters were also the

---

[16] This is a consequence of learning at the character-level, as words can only be associated to their context in terms of letter sequences instead of meaning or high-level abstractions. For example, even if the model is able to reproduce correct words, their position in a sentence might not be grammatically or semantically correct.

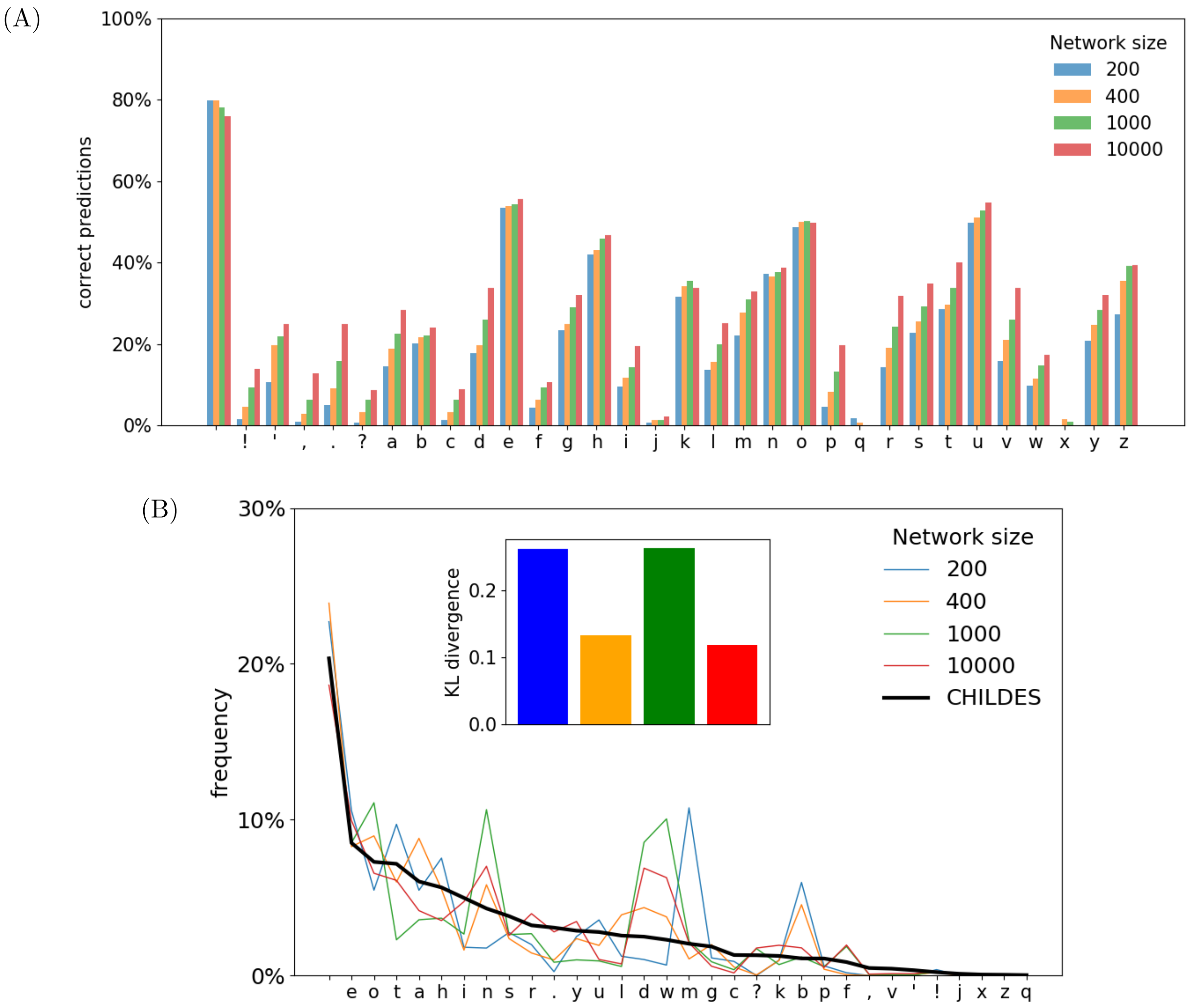


Figure 5.8: **Statistical learning in the CHILDES dataset.** (A) Correct predictions of the subsequent character in the input corpus, for various network sizes, including spaces (invisible first character to the left) and basic punctuation signs. Bar plots show the average scores for 10 independent simulations. (B) Character frequency in examples of spontaneous outputs ($T_{\text{auto}} = 50\,000$) for networks of different sizes, in comparison to the input CHILDES dataset. Inset shows the Kullback-Leibler (KL) divergence between the distribution of character frequencies in the example outputs and the CHILDES dataset input.

most frequent output characters on average), some letters were clearly over-represented in the spontaneous output. In fact, such a phenomenon pointed to the occurrence of short, repeating loops of words in the output, suggesting a limited capacity of the model in generating language-like sentences (Fig. 5.9). Although small loops were also present in the input dataset (a consequence of the repetition of short sentences during interactions), their predominant appearance in the output showed that, although $\mathrm{SORN_L}$ models recalled frequently repeated parts of the input, their dynamics did not generate enough variability in order to recreate input patterns indefinitely. This was likely a consequence of the deterministic nature of our model, and in particular, the deterministic output character choice by the readout layer. Nevertheless, at least for the tested scenarios, standard deterministic SORNs did not prove to be adequate models for the reproduction of the various temporal patterns contained in the CHILDES dataset and were only able to capture simple input statistics such as the overall letter count.

5.4.2 SORNs vs. "deep" recurrent neural networks

Given the complexity of real language input, deterministic SORNs were only capable of identifying and reproducing the temporal patterns in the CHILDES dataset to a limited extent. The performance on the prediction of the next letters did not exceed 50% even for the most frequent cases (such as the vowels "e" or "o"). Of course, optimal performance in this task is limited due the variable and sometimes unpredictable nature of spoken language, but a more detailed look at the autonomous output showed a limited reproduction of input patterns, with frequent recurrent *loops* (Fig. 5.9 — an example of SORN with $\mathrm{N^E} = 1\,000$ that reached the infinite loop of repeating characters "teddy bear. turn around!"). However, deep neural networks such as LSTMs have been shown to be able to learn a huge variety of patterns, even at the character level, and to produce autonomous outputs that resemble the input corpus (Karpathy, 2015). What are the main differences between SORNs and other recurrent neural networks typically used for natural language processing? In this final section, we summarize their main distinctive features and qualitatively compare their generative abilities, in order to better understand the mechanisms via which complex temporal encoding can be improved in self-organizing networks.

We chose to compare SORNs to a recurrent network of GRU units (see Appendix C for details on the GRU implementation and parameters). Differ-

SORN $N^E = 100$ $\sim 10^3$ variables	"ie he she see. ing the win the ba comy and he dour an mommy upand an mommy and sous down down down the s ople t you wan"
SORN $N^E = 1000$ $\sim 10^5$ variables	"what song should we sing now? let us go do teddy bear! you like this teddy bear. teddy bear. teddy bear. turn around! teddy bear. teddy bear. turn around! teddy bear."
GRU 100 units $\sim 10^5$ variables	"you wanna see if you can not have that. you are going to be right here! you are going to get you a big bird. here. oh, you are going to get you a second."
GRU 1000 units $\sim 10^6$ variables	"go to sleep mname mname boy, go to sleep my dear. we will have to sit up again so you can not have the birdie. there is a car. that is the way. there is the way and there is a baby"

Figure 5.9: **Sample spontaneous outputs for SORNs and GRUs.** Sample outputs for different recurrent networks after training on the first 500 000 characters of the CHILDES dataset, exemplifying the emergence of correct words (SORN, $N^E = 1\,000$), sentence patterns (GRU, 100 hidden units), and context dependent interactions (GRU, 1 000 hidden units). SORNs were trained for $T_{\text{plast}} = 500\,000$, $T_{\text{train}} = 30\,000$, and $T_{\text{auto}} = 20\,000$. GRUs were trained for 100 epochs, with sequence length of 100, batch size of 64, embedding dimension 256, and output temperature 0.3.

ently from SORNs, GRUs' learning rules relied on gradient descent methods rather than biologically inspired self-organization. In summary, the GRU model combined an input embedding representation, in which each input character was a multidimensional vector of fixed size, with a dense recurrent network of GRU units (Cho et al., 2014) and a fully connected readout layer, which predicted the next character based on a previous sequence of fixed length with a softmax output sampling (i.e., the output character was sampled from a softmax distribution with a given "temperature" parameter). The model was trained for a number of epochs, in which it received the whole dataset as input[17]. After training, its spontaneous output was gener-

---

[17] The reader might note that this is not the case for SORNs, which were trained only once on a subset of the CHILDES input corpus. Training SORNs on the whole dataset, however, was impractical due to computational time. Furthermore, there was no evidence that training SORNs on the whole dataset would lead to more realistic autonomous out-

ated via a similar input retro-fed process, which replaced the external input. It is easy to see the main reason why such a model was fairly different from reservoir computing networks and was expected to have higher performance when generating text: it contained a huge number of trainable variables and controllable hyperparameters. As a simple comparison, a SORN_L with 1 000 excitatory neurons contained approximately 10^5 trainable variables in its reservoir, while a GRU with the same number of hidden units included approximately 4×10^6 [18]. An additional important difference was the training procedure: while SORNs relied on a combination of self-organizational unsupervised mechanisms in the main reservoir, GRUs updated their weights in order to minimize a loss function in a supervised manner. There are little to no studies discussing how self-organization relates to gradient descent optimization in neural networks (but see section 6.2.2 for recent developments), and a formal comparison was beyond the scope of this short section. Finally, the GRU model drew its predicted output from a softmax distribution rather than always choosing the most probable character, which in practice resulted in a non-deterministic readout layer[19].

Instead of comparing the theoretical mechanisms responsible for the learning capacity in both SORNs and GRUs, which would require an extensive description of gradient descent based algorithms and the differential equations for the units' gates, we analyzed the models' empirical results on the CHILDES dataset by looking at their autonomously generated text (Fig. 5.9). For the sake of simplicity, we restricted our comparison to relatively small (or "shallow") GRU models, with a maximum of 1 000 hidden units in a single recurrent layer. The provided text samples illustrate typical outputs of models of different sizes, which were also observed in longer and independent simulations (not shown). First, small SORNs ($\text{N}^\text{E} = 100$) generated sentences that contained short, mostly incorrect words, with few exceptions. Bigger models ($\text{N}^\text{E} = 1\,000$), however, were already capable of generating

---

puts. A systematic comparison, therefore, was beyond our scope, and we suggest that future work should investigate the effects of the input corpus size for plasticity driven learning.

[18] Likewise, a GRU model with approximately 100 hidden units contained approximately the same number of trainable variables as a SORN with 1 000 excitatory units. This difference emerged from the multiple variables employed in each of the various gates of GRU units.

[19] Note that, qualitatively, the output softmax temperature parameter controlled the output variability, analogously to physical temperature. Temperatures close to zero resulted in less variability, while higher temperatures made the output more variable.

sentences with mostly correct words, but eventually reached recurrent loops of short sentences that could not be escaped and would be repeated indefinitely. In contrast, GRUs were already capable of generating correct words while avoiding loops for fairly small models with 100 hidden units. Although many of the generated sentences were very similar (such as the ones starting with "you are going to get" in the provided example), we already observed some variability in words, due to the softmax output. Finally, Larger GRUs (1 000 hidden units) showed variable outputs and were able to keep some semantic connections between sentences (which were, of course, learned from the input corpus).

It is insightful to compare the autonomous outputs of SORNs and GRUs with the same number of trainable variables (as discussed, 1 000 excitatory units for SORNs and 100 hidden units for GRUs), since learning capacity has been suggested to be a function of the number of units and different training methods in recurrent neural networks (Collins et al., 2016). As both outputs lacked context connections between sentences, the main contrast between them was the presence of repeating short loops in the SORN case, which were only temporary or absent in the output of GRUs. Given the GRU readout layer, however, this result was not surprising: in contrast to the deterministic SORN, the output prediction sampling from a softmax distribution with non-zero temperature could explain the higher variability and lack of repeating sequences of characters. In fact, the addition of softmax sampling (with a small temperature of 0.05) during the autonomous phase of the SORN led to more realistic language generation without loops, as exemplified by the quote at the beginning of this chapter, which is a sample output of the SORN after being trained with the first two sections (5.1 and 5.2)[20]. This result suggested two important conclusions. First, even though the SORN_L was not capable of generating enough variability in its autonomous outputs, it was able to learn many distinct temporal patterns from the input corpus. The output variability was limited, in fact, by the readout layer, which could easily be modified to generate more variable sentences. Second, self-organization combined with non-linear methods might lead to more realistic autonomous outputs, possibly even comparable to small recurrent neural networks of gated units trained via gradient descent.

---

[20] The language of the input corpus, in this case, was arguably more difficult to learn than the CHILDES dataset, reason why a realistic autonomous output required a network with more neurons. This result, nonetheless, was very well received by the author of this particular input text.

In summary, our brief comparison between SORNs and GRU models suggested that a deterministic self-organizing network is capable of encoding input temporal information, but requires a stochastic readout layer in order to autonomously generate similar output. Importantly, although self-organization in the SORN is inspired by biological plasticity mechanisms, the same is not true for the readout layer, which is a supervised multi-class classifier trained via logistic regression (see Chapter 3). Our results suggest that a more complex and possibly non-deterministic mechanism might be responsible for language acquisition and reproduction in biological neural circuits. Such mechanisms should work in addition to the combination of synaptic and homeostatic plasticity responsible for the encoding of input information. The combination of synaptic and homeostatic plasticity, furthermore, might assist spatio-temporal learning even for complex tasks that are commonly approached via supervised deep neural networks. Future work should reveal if and how unsupervised plasticity driven self-organization is able to assist current supervised neural network architectures and training methods in capturing different input temporal scales together, from character-level dependencies to the context of words and sentences.

5.5 Discussion

We have extended the analysis of criticality from the previous chapter in order to investigate its links to another brain function, the fading (or working) memory capacity, which measures how long past inputs can remain in a system's recurrent activity. This analysis was initially motivated by the maximum logarithmic memory scaling observed in reservoirs at the edge-of-chaos (Bertschinger and Natschläger, 2004), whose general relation to neuronal avalanches and self-organized criticality was unknown. Our results showed that a logarithmic scaling was indeed observed in the $\mathrm{SORN_L}$ when recalling distinct symbols from random past inputs, even though the same input was responsible for the breakdown of the power-law distributions of neuronal avalanches' sizes and durations in the $\mathrm{SORN_Z}$. When driven by a similar input from a counting task, however, the $\mathrm{SORN_L}$ has been shown to exhibit subcritical behavior via perturbation analysis (Lazar et al., 2009), resulting in an apparent contradiction to the observed critical memory scaling. Fortunately, this contradiction did not hold under careful analysis. First, edge-of-chaos dynamics do not generally co-occur with critical phenomena (Kanders et al., 2017a), and our results suggested SORN models exemplify this previous observation. Second, the critical memory scaling might have been achieved via self-organization even if the network did not display dynamics at the edge-of-chaos. Third, as discussed in the previous chapter, disentangling the input of sequence learning tasks and the model's dynamical state is particularly difficult. The lack of power-law scaling for neuronal avalanches in the case of structured, learnable inputs is not necessarily a proof of non-critical dynamics, but likely a consequence of the input temporal structure.

Given all those observations, what can we finally conclude about the relationship between criticality, neuronal avalanches, learning, and memory abilities? The connecting factor among all our results is the plasticity driven self-organization, which is capable of driving the models towards different dynamical states depending on the input conditions. This adaptation to input intensity and patterns has a parallel with different experimental setups (Priesemann et al., 2014): while neuronal avalanches appear only in spiking activity *in-vitro*, plasticity *in-vivo* could potentially act to improve functional properties such as the fading memory, while keeping the system away from criticality signatures due to the external drive. Interestingly, our results also suggested that memory curves are improved, but not maximized, at intermediate membrane noise levels, in which power-law scaling of neuronal avalanches occur[21]. Thus, the relationship among those mechanisms remains elusive, and we expect future studies to investigate the maximization of the fading memory capacity as a potential functional role of critical, or near critical, dynamics in the brain.

Whereas in models the link between processing capacity, memory and criticality has been widely investigated, experimental evidence is scarce, potentially because it is harder to obtain. The main challenge is to tune the experimental system precisely from sub- to supercritical states, ideally in a manner that does not impede its natural processing capacities in a given state. Additionally, the processing capacities, such as the fading memory, need to be quantified. The classical approach is to make use of pharmacological interventions to control the excitatory and inhibitory balance. For instance, the dynamic range obtained from local field potential (LFP) record-

---

[21] We emphasize the difference between improvement and maximization of the fading memory capacity. For instance, we have observed that a lognormal target firing rate distribution among excitatory neurons can also improve the SORN's learning and memory capacities, although it does not, a priori, maximize them.

ings *in-vitro* under electrical stimulation is maximized in an unperturbed system and diminished when excitation or inhibition is reduced pharmacologically (Shew et al., 2009). The same holds for the entropy of evoked patterns and for the mutual information between stimulus strength and response pattern (Shew et al., 2011; Shew and Plenz, 2013). Recent work has also shown that the mutual information between the past activity of two neurons and their future spiking increases with circuit maturation *in-vitro* (Wibral et al., 2017). Given that with maturation neural networks *in-vitro* approach a critical state (Levina and Priesemann, 2017), this clearly indicates that more information about the past can be read out in the future, as the network self-organizes towards a critical point. By characterizing contributions from the source neurons (Wibral et al., 2015), a complementary study showed that the relative contribution of synergy to mutual information increased, although the unique contributions from each source decreased during the first four weeks, indicating that the neural network developed information modification capabilities. In the fifth week, however, the redundant or shared contribution dominated, and information processing became highly similar across neurons, possibly due to a lack of external inputs (Wibral et al., 2017). Together, these studies showed that favorable information processing capabilities increased around criticality, and therefore agreed with our modeling results.

In the second part of the chapter, we tested the limits of learning and memory abilities in the $\mathrm{SORN_L}$ by employing more challenging temporal learning tasks, in the form of grammar learning, sentence generation, and language acquisition. It is important to mention that grammar learning has been attempted with SORN models before (Duarte et al., 2014), however with a different approach. Instead of learning temporal sequences of characters, the authors focused on simpler Reber grammars (Reber, 1967), which contained "sentences" composed of a small number of symbols and transition rules based on a predefined directed graph. The $\mathrm{SORN_Z}$ succeeded in distinguishing correct and incorrect input sentences (generated from a transition graph), although the prediction of future input symbols was not required. We have extended those results and shown that not only multiple real sentences can be correctly learned at the character level, but they can also be subsequently recalled and autonomously generated long after the external input was cut off. Additionally, *new* correct sentences could be generated, showing that output variability might arise even in deterministic models. Interestingly, neural variability akin to cortical spike recordings has been shown to also emerge in the same deterministic models of the SORN family (Hartmann et al., 2015), further emphasizing that plasticity driven self-organization in neural circuits can generate non-deterministic behavior.

Learning temporal patterns with recurrent networks is by no means a new topic, and simple random networks have long been introduced as a way of performing statistical learning and storing input structures (Elman, 1990). Curiously, today most natural language processing models encode information at the word-level via embedding representations (Mikolov et al., 2013), possibly due to the fact that words themselves do not need to be learned in these cases. This approach, however, is not able to detect the fine structure of symbols in a given sentence, relying instead on learned relationships between words, and has limited application for the investigation of how simple language structures are first acquired. We have attempted to describe language learning at the character-level by emulating statistical learning during language acquisition (Kuhl, 2004). Importantly, the sequences of letters in our models could be more realistically thought as phonemes, when spoken language was considered[22]. Unfortunately, SORN models were too simple to capture and reproduce all the complexity of spoken language, even when restricted to a dataset of interactions with infants. Nonetheless, we have shown that self-organization in the reservoir alone was able to learn some of the input character statistics. Interestingly, the lack of realistic autonomous output was a consequence of the SORN's linear readout layer, which could be combined with softmax sampling (typically employed in various machine learning architectures) in order to generate improved outputs. Such a result suggested that self-organization due to biologically inspired plasticity might be combined with other recurrent neural network architectures and learning algorithms for more efficient models, and we hope future biologically inspired network architectures can be designed or improved based on this combination.

Last, the fact that a combination of a deterministic neural network model with a simple stochastic output generation mechanism was able to give origin to realistic variability, either in the form of new correct sentences or "language-like" output, was remarkable. Although we did not explore the

---

[22] Precisely, the number of phonemes in a language is typically higher than the number of letters in its alphabet (English, for instance, has 44 phonemes). However, our results were robust regarding the alphabet size, and the neural input pool N^{U} for each letter could be increased or decreased as necessary to accommodate a higher or smaller number of distinct inputs, with negligible effects on our results.

full extent of possibilities and insights to be gained for language acquisition modeling, we hope that these results will pave the way and possibly inspire more detailed models of statistical language learning in infants. In particular, as our model mainly failed to capture long-range temporal patterns and correlations, we expect that hierarchical models that combine learning mechanisms acting at different levels might yield improved results and more realistic autonomous output in the future.

Chapter 6

Conclusion and outlook

> Good news, everyone!
>
> Professor Hubert J. Farnsworth

6.1 Summary

The brain's dynamical state is surprisingly stable (i.e., activity neither dies out or is repeatedly amplified) given the constant adaptation required by multiple input intensities and spatio-temporal patterns. This adaptation is regulated by various synaptic and homeostatic mechanisms acting at the neural level, which also play essential roles for information processing. Nonetheless, experimental evidence has suggested that this self-organization phenomenon might result in different dynamics depending on internal and external factors, while theoretical and computational models have repeatedly shown the existence of a critical state for neural networks, in which information processing is maximized. Therefore, it is straightforward to speculate that, at least under certain conditions, the brain is poised at criticality.

In this thesis, our goal was to investigate the role of self-organization in the emergence of criticality signatures in neural circuits and a potential connection between criticality and learning abilities in neural networks. This was achieved using a family of self-organizing recurrent neural network models (SORNs), which not only have been shown to outperform static reservoirs on sequence learning tasks but also combine biologically inspired plasticity mechanisms and have been able to reproduce neural variability and various features of cortical dynamics. In Chapter 4, we explored the mechanisms underlying the occurrence and maintenance of experimentally observed criticality signatures, focusing on neuronal avalanches with power-law distributed sizes and durations. Second, in Chapters 4 and 5, we observed that the same neural self-organizational mechanisms are responsible for the model's learning abilities and, perhaps surprisingly, may result in a logarithmic scaling of fading memory, so far only observed in networks at the edge-of-chaos. Finally, in Chapter 5, we explored applications of the critical-like improvement in memory and proposed a self-organizational model for language learning and generation.

6.1.1 Criticality in neural circuits

We have observed criticality signatures in the form of power-law distributed bursts of neural activity in the spontaneous activity of the $\mathrm{SORN_Z}$ model. These signatures resulted from plasticity driven self-organization and resembled the neuronal avalanches detected in multiple experimental setups, including their transient break down under random external input and total absence under simple inputs containing learnable temporal patterns. Importantly, not only we found the power-laws to be input dependent, but they required a suitable level of membrane noise to occur, suggesting a potential experimentally testable control parameter. The contrast between the presence of neuronal avalanches in the model's spontaneous activity and their absence under simple structured input has a straightforward analogy to spiking activity *in-vitro* and *in-vivo*, and we have shown that the *same* combination of Hebbian and homeostatic plasticity can account for both cases, leading the system towards to and also away from criticality depending on the input. We highlight, again, that such a self-organization process could be highly advantageous for the brain: while a near critical dynamical state takes advantage of various information processing benefits, it avoids dangerous epileptic regimes.

Ultimately, does the critical brain hypothesis hold true and the brain is indeed poised at a phase transition state? A short answer is maybe, but likely not always. A slightly longer answer is that seemingly incompatible experimental measurements on criticality signatures might, in fact, be reconciled when self-organization and local input levels are considered. The results on SORN models are less speculative: criticality signatures appear in the $\mathrm{SORN_Z}$ spontaneous activity, while the $\mathrm{SORN_L}$ is slightly subcritical when performing learning tasks[1]. The presence of power-law distributed neuronal avalanches, however, was not sufficient to prove that the SORN_Z belongs to a class of self-organized criticality (SOC) models. In fact, the power-law exponents were different from the ones typically observed in experiments and branching processes, which reflected the emergence of more complex dynamics after self-organization. Given the effects of the noise level, we also note that the structure of particular inputs might generate apparent criticality in non-critical systems and vice-versa. Therefore, the input drive should be disentangled from the internal model dynamics when drawing further conclusions about criticality. Last, the term "criticality" itself is used in the literature to refer to at least two generally distinct dynamics, SOC phenomena and phase transitions from ordered to chaotic states. We have shown that SORNs are biologically inspired examples of complex models in which both do not, necessarily, co-occur.

6.1.2 Criticality meets learning and memory

As we have discussed in Chapter 4, external input, including the drive from other brain areas, is crucial for the local emergence of criticality signatures and possibly critical dynamics. However, linking criticality or the deviation from it to the performance in a particular task is challenging. For example, in this thesis we have shown apparently contradictory results: neuronal avalanches coincide with an increase in the fading memory capacity of the SORN_Z model, but they are not connected to maximal performance in simple learning tasks with structured input. While the relation between fading memory and performance in various tasks has not yet been clarified, this discrepancy can be explained by the nature of the learning task and the readout training procedure from reservoir computing. Although the addition of membrane noise level results in an increased number of internal representational states, suggested by the increase in the fading memory capacity, such effect cannot be fully exploited by a linear supervised readout layer, as the noisy network activity increases the error of the classifier. Additionally, as the power-law distributions seem to require unstructured input, their absence under repeating input sequences was not surprising, since the network can still achieve an internal structure that is beneficial for pattern learning.

---

[1] Recall that the differences between these models are two plasticity mechanisms, inhibitory spike-timing-dependent plasticity (iSTDP) and structural plasticity (SP), and the presence of neuronal membrane noise.

These results can be interpreted as a deviation from avalanche criticality due to structured input, while other network information processing abilities, including the fading memory, remain unaffected. Therefore, criticality signatures might still indicate a favorable, although not unique, dynamical regime for learning in recurrent networks.

In the last part of our study (Chapter 5), we have shown that self-organization in reservoirs ($\mathrm{SORN_L}$) leads to a logarithmic scaling of fading memory which only occurs for static reservoirs at the edge-of-chaos. This result has two main consequences for reservoir computing. First, self-organization might be responsible for maintaining a critical-like scaling even without poising the model at criticality. Second, an improved fading memory capacity in reservoirs indicates they are able to learn more complex temporal tasks. Although our results on language learning fall short of the current state-of-the-art deep learning frameworks in terms of performance, self-organization via biologically plausible Hebbian and homeostatic plasticity mechanisms might lead to future insights in statistical models of language acquisition. In particular, the successful generation of novel, correct sentences following predefined grammar rules and containing words from artificially built dictionaries suggests that self-organization might play a key role in how brain networks are able to encode complex temporal patterns.

6.2 Outlook

Our results suggested a few possible directions for follow-up studies, particularly regarding the role of criticality for brain function and the mechanisms governing the adaptation to different input conditions. Additionally, given the combination of self-organization and reservoir computing, or more generally machine learning, the implications of our studies might be relevant beyond the field of computational neuroscience.

6.2.1 Criticality and neural circuits

The debate about criticality in the brain has evolved since the observation of its first experimental evidence (Beggs and Plenz, 2003) but remains far from settled. Our results suggest that future theoretical work should focus on clarifying the relationship between power-law scaling (avalanche criticality) and “true” second-order phase transitions (as in sandpile models and branching processes) in various classes of dynamical systems. Although multiple models assume they co-occur, SORNs and other complex systems (Marković and Gros, 2014; Kanders and Stoop, 2016) are examples in which this relationship is not trivial and simple analogies to SOC might fail. In particular, our results support the idea that criticality analyses in brain circuits should always take into consideration the system's external drive, since biological self-organization mechanisms, including synaptic plasticity, seem to be able to sustain distinct dynamical states. This is potentially a consequence of the lack of separation of time scales in neural networks, and developing theoretical and experimental approaches to disentangle internally and externally driven dynamics are important future challenges. Interestingly, similar network adaptation has recently been observed at the whole-brain level (Ponce-Alvarez et al., 2018), in which external inputs are counterbalanced by plasticity action resulting in a repeated return to a phase transition point (such counterbalance by plasticity action can be achieved via equilibrium of excitatory and inhibitory synapses, as shown by recent work (Agrawal et al., 2018)). This experimentally observed process is comparable to self-organization and the emergence of criticality signatures we described in the SORN's spontaneous activity.

As for our modeling approach, obvious next steps are the investigation of experimentally motivated changes in the SORN's synaptic and homeostatic plasticity mechanisms and their effects on the system's dynamical state. Regarding synaptic plasticity, extensions of the STDP rule in the form of reward modulation have already been introduced in previous works (Savin and Triesch, 2014; Aswolinskiy and Pipa, 2015), suggesting that self-organizing neural networks might be used as models of reward-dependent learning under the theory of three-factor learning rules (Frémaux and Gerstner, 2016). Our own preliminary experiments have shown that reward dependence can be effectively combined with Hebbian learning for simple reinforcement learning tasks, although any link to critical dynamics, or even criticality signatures, remains unexplored. Self-organization via homeostatic plasticity, in addition, can be modified to account for the lognormal distributions of neuronal target firing rates observed in cortical tissues. Our exploratory tests, in fact, have shown that non-uniform firing rates slightly increased the SORN's performance in some of the learning tasks we studied, which should be further investigated by future work. Finally, a question that remains to be answered is the scalability of our results: how does self-organization affect much larger networks, with size and topology comparable to real neural circuits? The fact that critical dynamical behavior might emerge in hierarchical network architectures even when underlying processes are not critical (Friedman and Landsberg, 2013) suggests that hierarchical SORN-like models are an interesting possibility to study self-organization in large scale brain activity, and we believe our current results might provide important insights for the design of future large scale neural networks.

6.2.2 Self-organization and computing

We have provided empirical evidence that self-organization and criticality signatures are linked to learning and improved memory scaling in recurrent neural networks. However, we have left a formal description of the relationship between these capacities for future studies. As the scaling of these properties is arguably task dependent, a mathematical formulation of particular classes of tasks that may be optimized at critical or near critical dynamical regimes could improve the fine tuning of parameters of neural networks constructed to excel at those tasks. We have suggested language acquisition and generation as possible applications of deterministic, self-organizing reservoirs, but other applications for supervised and unsupervised learning could be investigated in the future. For example, due to the combination of various temporal patterns in language, hierarchical self-organizing networks could be particularly effective and greatly improve the results we obtained with SORNs. Furthermore, SOC analogies have recently been proposed for complex optimization problems in deep learning such as unsupervised image segmentation, in which costly parameter search methods could be replaced by self-regulation rules (Hoffmann and Payton, 2018). More generally, maps between self-organization phenomena (and renormalization theory) and modern deep learning frameworks have been suggested (Mehta and Schwab, 2014; Martin and Mahoney, 2018), and the possibility that biological plasticity mechanisms can improve standard gradient descent based techniques provides an interesting new topic of investigation that lies at the intersection of computational neuroscience, machine learning, and physics. In that sense, our description of self-organization towards criticality in small "toy" networks might not only shed light on important properties underlying biologically inspired learning models but also provide insights for the setup of more effective architectures of artificial neural networks.

Appendix A

Model implementation in python

During the course of this thesis, we developed two distinct implementations of SORN models. The initial implementation was the SORN repository in python 2.7, which was based on a former repository from previous studies (Hartmann et al., 2015, 2016). This version was employed to generate and analyze the results in Chapter 4, and the exact code to reproduce our analyses and figures has been made publicly available on *github*, as part of our publication (Del Papa et al., 2017). The maintenance of this old version, however, has become difficult due to conflicts in dependencies, multiple packages versions, and overall lack of backwards compatibility. Therefore, a second implementation and a new repository, SORN-V2, were created, including an update of the model to python 3.6 and more organized installation guides and reproducibility instructions combined with improved general software development practices. This new version was employed to generate the results, analyses, and figures in Chapter 5, as well as the follow-up book chapter (Del Papa et al., 2019). Both implementations are compatible and yielded equivalent results for all our tests, including the replication of the original SORN_L (Lazar et al., 2009) and the extended SORN_Z (Zheng et al., 2013) models. This is to be expected as they share some of the core classes and functions. In order to avoid confusion, we describe here the main differences between both implementations and provide basic instructions for future use, although we strongly recommend interested users to choose SORN-V2, as the old SORN repository is no longer maintained and will likely create various conflicts with different versions of old python packages.

A.1 SORN

Repository: `https://github.com/delpapa/SORN`

The main SORN dynamics are relatively easy to implement in most programming languages, as the update steps are a result of simple independent learning rules (Chapter 3). In order to properly manage experiments, test different sets of parameters, and plot desired results with relative ease, each experiment was organized in a separate module and kept apart from core functions that are shared by most simulations (e.g., the functions regulating plasticity rules or the readout training). This framework was already in place from previous implementations (Hartmann et al., 2015, 2016) and was initially kept the same for compatibility reasons. A new experiment module to compute neuronal avalanches was implemented in addition to other former experiments, with respective configuration files and plot scripts for the figures shown in Chapter 4.

Due to the nature of the neuronal avalanche analyses, which typically required distributions of thousands of events, independent simulations of millions of time steps were routinely required. A single SORN_Z simulation, for example, required a few hours in this framework (CPU implementation), which was already optimized by taking advantage of sparse matrices (the time bottleneck, as expected, was in the matrix multiplication operations required for some plasticity mechanisms and readout training). This large computational time cost demanded the implementation of a two-stage process: the main SORN simulation runs separately from the data analysis and plots, while intermediary results (for example, the total SORN excitatory activity per time step) were stored in disk (naturally, these intermediary results were not uploaded in the repository, and any new user would need to generate them again). Importantly, on top of the original implementation, the *powerlaw* package (Alstott et al., 2014) was necessary in order to plot power-laws and estimate exponents of various candidate distributions for the neuronal avalanches sizes, durations, and exponent ratios. The SORN simulation and experiments, finally, relied on additional standard python modules (*numpy*, *scipy*, *matplotlib*, among others).

A.2 SORN-V2

Repository: `https://github.com/delpapa/SORN_V2`

This follow-up repository includes the implementation of all experiments presented in this thesis, namely Counting Task, Random Sequence Task, Neuronal Avalanches, Memory Avalanches (which combines the neuronal avalanches and fading memory capacity analyses), Grammar Learning Tasks, and, finally, Language Learning Tasks. Experiments have been made into individual modules that can be run independently with different combinations of parameters and active plasticity mechanisms, although they share the common update methods and utility functions. Each experiment module refers to a different section of this thesis. The Counting and Random Sequence Tasks, as well as the Neuronal Avalanches, have been employed to validate the results from Chapter 4. The Random Sequence Task and the Memory Avalanches have been used to obtain the results regarding the fading memory capacity in Chapter 5. Finally, the Grammar and Language Learning Tasks have yielded the results from the second half of Chapter 5, including the generative language experiments.

The Counting Task was able to validate the performance results from the original SORN_L study (Lazar et al., 2009), while the Neuronal Avalanches experiment replicated results from SORN_Z (Zheng et al., 2013), including the lognormal distribution of synaptic efficacies and the evolution of the number of active excitatory synapses over time. From a practical perspective, each experiment module was divided into three parts for easier modular implementation: a configuration file, where all the simulation parameters are given, an experiment description script, in which particular experimental instructions are stored, and a source script containing all the information about the particular input source. Additionally, the majority of plot scripts have also been provided, allowing for the reproduction of this thesis' main figures (however, as in the previous repository, the main SORN simulation and analyses must be run separately and the intermediate results must be stored locally). The SORN-V2 implementation relied on the updated versions of the packages from the previous SORN repository, but now appropriate instructions for environment setting have been provided in order to avoid version conflicts in different machine setups. Further instructions and command line commands are provided in the repository's public web page, and interested users should be able to easily run a single simulation of the model.

Appendix B

Fitting power-laws with maximum likelihood estimators

As we have seen through this thesis, power-laws form the bulk of experimental evidence supporting the critical brain hypothesis. Additionally, we investigated and showed multiple power-law distributions in the SORN's activity, on which most of our conclusions are based. Given that the process of fitting power-laws is, in contrast to the common intuition, not straightforward, we describe here the method employed during this thesis to fit power-laws (Clauset et al., 2009).

Power-laws typically attract attention due to their mathematical properties and their presence in myriad natural and artificial systems, but their fit is particularly difficult due to uneven variations in the distribution's tail (Goldstein et al., 2004; Clauset et al., 2009). This noise in the tail can be both the result of the statistical nature of the system or even rounding errors, and already requires a specific logarithmic binning process, particularly for discrete data (Milojević, 2010). Mathematically, a variable x is said to follow a power-law if it is drawn from a distribution of the form:

$$p(x) \propto x^{-\alpha} \tag{B.1}$$

in which $\alpha > 0$ is the exponent (also known as slope or scaling parameter). In practice, power-laws appear above a $x_{\min} > 0$ value, as the distribution diverges for $x \to 0$. With the additional constraint of $\alpha > 1$, normalizing the

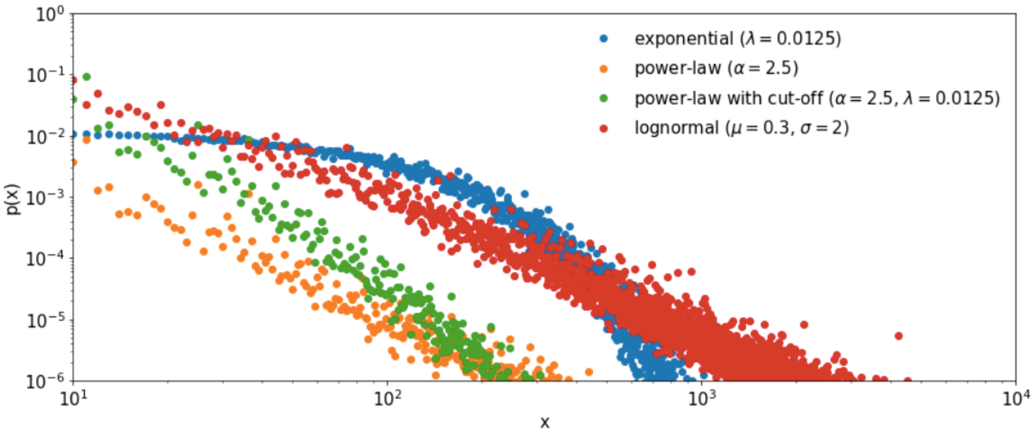


Figure B.1: **Power-law-like distributions in a log-log plot.** Comparison between randomly sampled distributions. Under a simple visual inspection, exponentials, lognormals, and power-laws look alike and can, in principle, resemble one another. More rigorous statistical tests such as MLE combined with goodness of fit are required in order to correctly classify the distributions.

distribution results in:

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \tag{B.2}$$

in which $\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha}$ is the generalized zeta function. The problem of fitting such a distribution to experimental data can be roughly divided into two steps. First, the best-fit parameters, α and $x_{\min}$, have to be appropriately calculated. Second, a goodness of fit parameter should be estimated, comparing how good a power-law fit is to other similar functions. Common choices of similar functions are exponentials, stretched exponentials, power-laws with exponential cut-offs, and lognormal distributions (Clauset et al., 2009), all of which look alike under a simple visual inspection of a log-log plot or histogram (see Fig. B.1 for a few examples).

The problem of fitting parameters is more complex than it seems at first sight. As power-laws typically have huge statistical fluctuations at the distributions' tails, a simple linear regression fit (for example, least squares regression) is biased and can lead to spurious exponents (Goldstein et al., 2004). In fact, most of the important information about the distribution is contained in the first points, and giving them more weight in the fit can already result in smaller variances. The most precise method to derive exponents in this case, for an assumed power-law distribution, is the maximal likelihood estimate (MLE; Clauset et al. (2009); for a more general and less formal tutorial on maximum likelihood with simpler examples, see Myung (2003)). Assuming that $x_{\min}$ is given, the probability, or likelihood, that n observations of continuous data x were drawn from a "pure" power-law distribution with exponent α is:

$$p(x|\alpha) = \prod_{i=1}^{n} \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \tag{B.3}$$

Thus, we should maximize this likelihood with respect to α in order to find to best exponent fit α^*. Alternatively, maximizing the logarithm of this distribution:

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln[p(x|\alpha)] = \frac{\partial}{\partial\alpha}\sum_{i=1}^{n}\left[\ln(\alpha-1) - \ln(x_{\min}) - \alpha\ln\left(\frac{x_i}{x_{\min}}\right)\right] = \\ n\ln(\alpha-1) - n\ln x_{\min} - \alpha\ln\left(\frac{x_i}{x_{\min}}\right) \overset{!}{=} 0\end{aligned} \tag{B.4}$$

results in the MLE for the exponent:

$$\alpha^* = 1 + n\left[\sum_{i=1}^{n}\ln\left(\frac{x_i}{x_{\min}}\right)\right]^{-1} \tag{B.5}$$

in which convergence happens in the limit of large n. The case of discrete data x is, however, more complicated and has no general closed form (Goldstein et al., 2004). Following Clauset et al. (2009), the MLE for the discrete case can be approximated with high accuracy ($\approx 1\%$) by:

$$\alpha^* \approx 1 + n\left[\sum_{i=1}^{n}\ln\left(\frac{x_i}{x_{\min} - 0.5}\right)\right]^{-1} \tag{B.6}$$

given that $x_{\min} > 6$. For the purposes of this thesis, we tested both discrete and continuous exponent estimates for the neuronal avalanche data resulting from our SORN simulations (which are discrete). As suggested by the original work (Clauset et al., 2009), we opted for the discrete approximation, as it provided higher accuracy, especially for $n > 50$ (recall that, in our data, we had thousands of events per simulation, resulting in hundreds of thousand data points per distribution).

In order to estimate the second fit parameter, $x_{\min}$, different methods can be employed, including a visual estimation in a log-log plot (Clauset et al., 2009). For simplicity, we chose the discrete value which minimized the distance of our experimental distribution to a pure power-law, by testing the exponents and errors for each possible $x_{\min}$ value. This approach, if not carefully done, can introduce biases towards power-laws, which is an additional reason to compare the goodness of fit between power-laws and other alternative distributions.

The goodness of fit can be considered an estimate of how plausible a power-law distribution is to the given data. In practice, estimating the best-fit parameters may result in the best power-law fit, but this procedure provides no information about the plausibility of a power-law distribution for the data. Ideally, a goodness of fit measure should detect if deviations from a pure power-law are purely statistical or actually suggest that the given dataset follows another distribution. Naturally, it is always possible to overfit any dataset with a distribution defined by enough parameters, while it is impossible to compare *every* possible distribution and choose the best fit. Therefore, some prior knowledge about the dataset is always required, and in practice, typical distributions compared to power-laws are heavy-tailed distributions with only one fitted parameter. Any two distributions can be compared using a loglikelihood ratio test (Clauset et al., 2009), among other approaches. This test estimates the loglikelihood ratio R between two distributions at data points x_i. Assuming the test is to be made to compare distribution $p_1(x)$ and $p_2(x)$, R is given by:

$$R = \ln \prod_{i=1}^{n} \frac{p_1(x_i)}{p_2(x_i)} = \sum_{i=1}^{n} [\ln(p_1(x_i)) - \ln(p_2(x_i))] \tag{B.7}$$

in which both logarithms are, in practice, the loglikelihood for a single measure x_i. For the limit of large n, R is positive is case p_1 is more likely and negative in case p_2 is more likely, and a significance p can be estimated based on the variance of the distribution of R[1]. In practice, if p is small enough ($p < 0.1$), it can be shown that R is unlikely to be a chance result and can be

---

[1] Of course, for the sake of brevity, we are glossing over many details on the calculation of R. See Clauset et al. (2009) and references within for those details.

trusted when choosing the more accurate distribution to represent the data.

Fortunately, the estimation of α, x_{min}, R and p can be done using automated python packages developed specifically for this purpose and based on MLE, using the same procedure briefly described here — namely the *powerlaw* module (Alstott et al., 2014). During our work, we have first employed our own functions to estimate those parameters but switched for the pre-implemented ones and *powerlaw* for consistency and computational speed (as far as we have tested, the estimated exponents were similar when using our functions or the package). Therefore, all the results shown in this thesis have been plotted and estimated using this particular python package, and the current SORN implementation depends on it for the avalanche analysis.

Appendix C

Grammar and language learning

In Chapter 5, we studied the SORN's grammar and language learning abilities by employing artificially constructed dictionaries, an open dataset containing transcripts of language acquisition in infants, and finally implemented a simple "deep" recurrent neural network composed of Gated Recurrent Units (GRUs; Cho et al. (2014)). Here we provide additional information about each of those auxiliary tools.

C.1 Predefined dictionaries

In order to create simple, language-like inputs composed of sequences of symbols with various time dependencies, we created 3 distinct dictionaries to serve as input to the SORN: the " fox drinks tea." (FDT), the extended FDT (eFDT), and the singular/plural sentences (SinPlu). Each contained a different number of words and grammar rules that not only made the learning tasks increasingly difficult but also captured some properties of real language.

FDT This first dictionary contained sentences of the form "[*subject*] [*verb*] [*object*].", including spaces and the final stop, and from which words in each category were randomly drawn from a fixed list. This list contained 8 different subjects, 2 verbs, and 8 objects, with the constraint that each verb could only be followed by 4 exclusive objects (Fig. C.1A), resulting in a total of 64 possible sentences with an alphabet size $\mathrm{U_A} = 25$ (i.e., containing almost

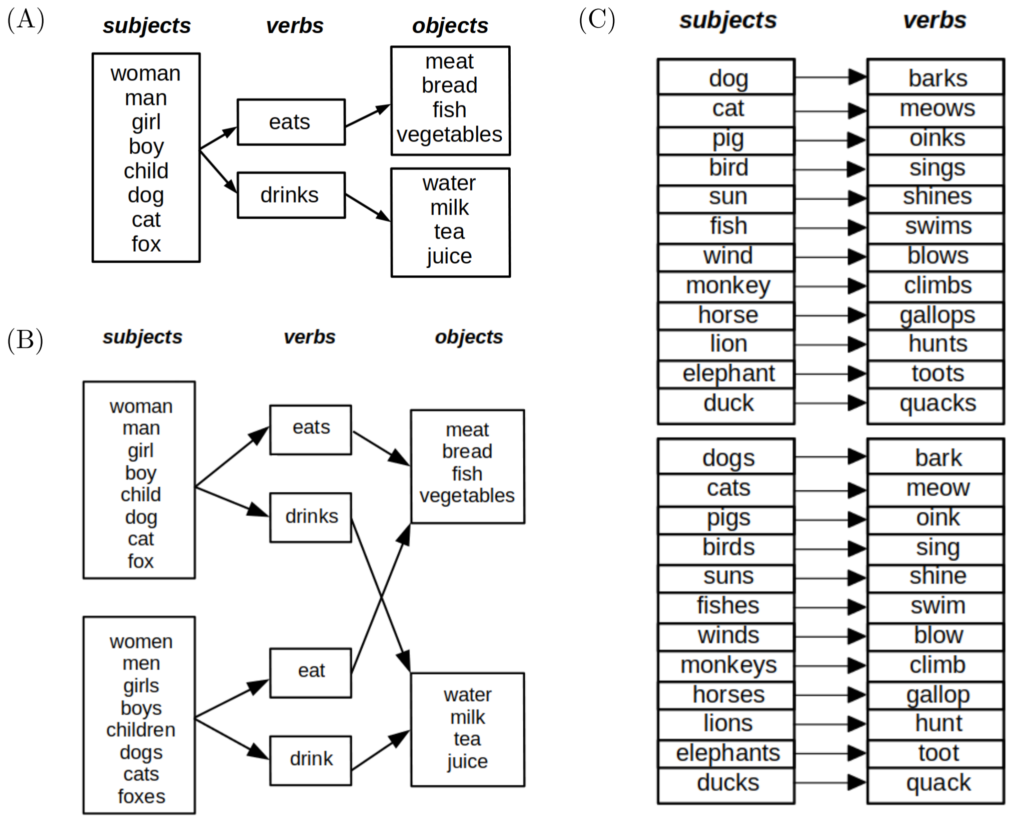


Figure C.1: **Artificial dictionaries for the grammar learning tasks.** (A) Fox Drinks Tea dictionary (FDT). Sentences were constructed by random selection of *subjects*, *verbs*, and *objects*, according to the arrows: each verb could only be followed by half of the objects. (B) Extended FDT dictionary (eFDT). Sentences constructed by this dictionary included the singular and plural forms of the ones in the FDT, duplicating the number of its *subjects* and *verbs*. (C) Singular/plural dictionary (SinPlu). Sentences were constructed via the random selection of *subject-verb* pairs from 12 possibilities, which could be either in their singular or plural forms (thus, 24 distinct possible distinct sentences). *Subjects* and *verbs* were associated according to the arrows.

all lower case letters of the English alphabet). In order to construct an input corpus, sentences were randomly selected, with equal probability, and concatenated while preserving the punctuation. For simulations in which a number of randomly selected sentences were removed from the input, their probability of selection was set to 0. As an example, an input fragment could be composed of the sentences " fox drinks tea. cat eats meat. dog eats vegetables.".

eFDT As an extension of the FDT dictionary, eFDT contained sentences according to the same grammar patterns and verb-object constraints, but also including plural forms of subjects and their respective verbs (Fig. C.1B). Overall, this dictionary contained 16 subjects, 4 verbs, and 8 objects, for a total of 128 possible distinct sentences (alphabet size remained $\mathrm{U_A} = 25$).

SinPlu In analogy to the parsing of plural indicator tokens, commonly observed in statistical learning models of natural language (Willits et al., 2009), this dictionary was composed of 12 *subject-verb* pairs ($\mathrm{U_A} = 24$), which could be either in the singular or plural form (Fig C.1C). The main difference between singular and plural forms was the position of the "s" character or plural token: for singular sentences, it should be placed after the *verb* and not after the *subject*, while for plural sentences the exact opposite should take place. The input corpus constructed from the SinPlu dictionary concatenated randomly chosen sentences, with equal probability, independently of their singular or plural state. For example, a fragment could read " dogs bark. duck quacks. dog barks. cat meows."

C.2 The CHILDES dataset

To address a more realistic language learning framework, we also trained SORNs on the CHILDES corpus, an open database containing transcripts of interactions with infants in various types of situations (MacWhinney, 2000). The CHILDES database contains multiple short transcripts of activities such as reading, playing, and eating, and provides a relatively realistic overview of vocabulary and grammar structures to which infants are commonly exposed. Importantly, this database reproduces transcripts of native American English spoken language, and thus contains more variable patterns and more flexible structures that are not commonly found in written texts. In our simulations, we employed a raw unparsed version of the dataset which included 4568 short transcripts of different interactions involving infants and children up to 72 months of age. This was an expanded version of the dataset employed by a previous language acquisition study (Huebner and Willits, 2018), whose authors kindly provided their corpus text files. The CHILDES does not, of course, reproduce the full range of language interactions infants typically are subject to, but should be viewed as an approximation of the semantic and grammatical structures that occur during various stages of development.

The raw corpus contained capital and lower case letters, numbers, various punctuation marks[1] and onomatopoeias, and almost no pre-processing, except for the substitution of female, male, and genderless names for “FNAME”, “MNAME”, and “ANAME”, respectively. Overall, the corpus had an alphabet size of $\mathrm{U_A} = 59$ (including capital and lower case letters, spaces and punctuation), and contained approximately 1 million sentences, composed of roughly 6.5 million words (or 32 million characters). In order to make the corpus less variable, and therefore easier for the SORN to learn, we converted all letters to lower case and removed all numbers and punctuation characters with fewer than 1000 occurrences (the remaining ones were: “ . ”, “ ! ”, “ ? ”, “ , ”, and “ ’ ”). Such data processing reduced the alphabet size to $\mathrm{U_A} = 32$ while maintaining the corpus size roughly unaltered.

C.3 A simple GRU network

In the end of Chapter 5, we presented text generation results obtained from a character level based recurrent neural network with gated recurrent units (GRUs; Cho et al. (2014)). Differently from reservoir computing, this model consisted of a “deep” network with three layers: embedding, GRU, and dense. Its implementation followed a text generation example in *tensorflow* (TensorFlow, 2018), based on a famous experiment with deep LSTM models and Shakespearian text (Karpathy, 2015). As with SORNs, the GRU model was initially trained on a prediction task, by estimating which was the next most probable character to appear in a sequence given the recent character his-

---

[1] Punctuation marks were originally provided as tokens, e.g. “PERIOD” or “QUESTION” for full stops and questions marks, respectively. These tokens are useful for word-level learning models, which treat each of them as separate words. In our character-level learning model, they were substituted for the respective punctuation symbol, in order to maintain each token as a single character.

tory. Without going into details of the specific model implementation[2], we qualitatively describe here the behavior and functions of each layer, in order to contrast the learning rules with the self-organization mechanisms of the SORN models.

The first difference between GRUs and reservoirs is the use of an embedding layer, which expands the representation of each character into a high-dimensional space before using it as input to the main recurrent network. This high-dimensional representation may also evolve over time, and characters that appear nearby frequently tend to have a lower "distance" on a high-dimensional Euclidian space, thus facilitating the network training. This method is inspired by a word-level representation algorithm, known as *word2vec* (Mikolov et al., 2013), and adds the embedding dimension as a hyperparameter to the GRU model. The output of this embedding layer is received by a recurrent network, roughly analogous to the main SORN reservoir, but with two important differences. First, GRU units implement complex non-linear functions with multiple internal variables, taking advantage of "gates" that learn when to store or release the units' input information. Second, each network update is performed at the same time for a sequence of characters instead of a single one. In practice, that means the units' state is updated (via backpropagation through time) after using a sequence of fixed length (typically 100 characters) to predict the probability for each next character. Finally, the probability of each output character is predicted via a dense, fully connected layer, which plays a similar role to the readout layer in reservoir computing.

After the training phase, in which the GRU model is presented with the whole text dataset for a number of epochs, its output is fed back as input, in an autonomous phase. This phase, however, is also different from our SORNs, since GRUs explicitly use a sequence of fixed length (again, typically around 100 characters) as input to predict each next character, instead of relying only on the internal temporal encoding. In practice, this means that, at every time step, a number of characters (in their embedded form) are used as input, in contrast to only a single one in SORNs. Although this procedure clearly improves the encoding of long-range temporal information,

---

[2] The interested reader can look up the source code from a side project at `https://github.com/delpapa/TweetGen`, where we have used the same model to generate new tweets based on the statistical patterns of the twitter history of a given user. It is straightforward to change the input source to, for example, the CHILDES dataset or any other desired text.

we highlight that this method explicitly relies on multiple past time steps, and any single output character is presented to the model multiple times during the subsequent time steps.

Last, we note that even a relatively small GRU model has millions of variables. For example, a model trained on the English alphabet (including punctuation, upper and lower case characters — an alphabet size of $\mathrm{U}_\mathrm{A} = 65$) with an embedding dimension of 256, sequence length of 100, batch size of 64, and 1024 GRU units has approximately 4 million variables, most of which in its recurrent units. For the purposes of this thesis, we followed the training procedure suggested in TensorFlow (2018), employing the Adam optimizer for backpropagation during 100 epochs. As in every machine learning model, we emphasize that our results are highly dependent on the particular parameter tuning and model architecture, and exploring their effects for language learning and autonomous generation at character-level was out of our scope.

List of Figures

2.1 **Examples of self-organization and emergence in diverse natural systems.** (A) Color pattern emergence in adult ocellated lizards. (B) Sand dune ripples formed by wind action. (C) Large-scale emergent brain networks. 11

2.2 **Three phases of the 2D Ising model.** Snapshots of two-state spin configurations for three different temperatures (subcritical, critical, and supercritical). 13

2.3 **Scaling in the Bak–Tang–Wiesenfeld sandpile model.** (A) Normalized avalanche size distribution. (B) Normalized duration distribution. (C) Power-law scaling of average avalanche size. (D) Snapshot of emergent symmetrical patterns. 17

2.4 **Avalanches in stochastic branching processes.** (A) Normalized avalanche size distribution. (B) Analogous distributions for avalanche duration. (C) Scaling of the average avalanche size. (D) Sample activity distribution for a system at criticality. 21

2.5 **Self-organization mechanisms in neural circuits.** (A) Spike-timing-dependent plasticity (Bi and Poo, 1998). (B) Inhibitory spike-timing-dependent plasticity (Haas et al., 2006). (C) Synaptic scaling (Watt and Desai, 2010). (D) Structural plasticity (De Paola et al., 2006). (E) Intrinsic plasticity (Desai et al., 1999). 26

3.1 **Artificial neurons.** Representation of the McCulloch and Pitts processing unit. 37

3.2 **Single and Multi-Layer Perceptrons.** (A) A single layer perceptron. (B) A Multi-Layer Perceptron (MLP). 38

3.3 **The Self-Organizing Recurrent Neural network model (SORN).** (A) Diagram of the SORN. (B) Diagrams of synaptic plasticity mechanisms 44

4.1 **Neuronal avalanches in cortical circuits, from Beggs and Plenz (2003).** (A) Power-law distributions for avalanche sizes, based on the number of activated electrodes. (B) Probability distribution of avalanche size based on summed LFPs. (C) Time binning procedure for avalanches. 61
4.2 **Experimental evidence of criticality signatures in various circuits.** (A), (B) (Pasquale et al., 2008). (C), (D), (E), (F), (G) (Friedman et al., 2012) 63
4.3 **Avalanches *in-vivo*.** (A) (Tagliazucchi et al., 2012). (B) (Petermann et al., 2009). (C) (Scott et al., 2014). (D), (E), (F) (Priesemann et al., 2014). 65
4.4 **Adaptation tunes cortical dynamics back towards a critical dynamics after short transient. (Clawson et al., 2017).** (A) Experimental setup for *ex-vivo* recording in the turtle visual cortex. (B). 67
4.5 **Benefits of critical dynamics for information processing.** (A) (Kinouchi and Copelli, 2006): dynamical range.(B), (C): (Shew et al., 2011): mutual information and information capacity. (D) (Haldeman and Beggs, 2005): information capacity. 73
4.6 **Self-organization phases and neuronal avalanches.** (A) Fraction of active connections. (B) Neuronal avalanche event definition via activity threshold. 80
4.7 **Neuronal avalanches in the SORN's stable phase.** (A), (B) Normalized distributions of size S and duration T. (C), (D) Scaling of avalanches for networks of increasing size. (E) Avalanche average size as a function of duration. (F) Power-law scale range, up until the cut-off point, as a function of network size. 83
4.8 **Robustness of choice of activity threshold.** (A) Activity distribution. (B) Avalanche size distribution for different thresholds. 85
4.9 **Alternative avalanche definition.** Example of avalanche size distribution. 86

4.10 **Effects of logarithmic binning on the avalanche distributions.** 87
4.11 **SORN_Z with frozen plasticity.** (A) Distributions for random networks and SORN with full frozen plasticity. (B), (C), (D) Distributions for combinations of frozen plasticity mechanisms 88
4.12 **Noise level gives rise to criticality signatures.** Top row: activity distributions for distinct noise sources and levels. Bottom row: distributions of avalanche sizes for the same noise forms. 91
4.13 **Raster plots for distinct noise levels.** 92
4.14 **SORN_Z readaptation after random external input onset**. Size (A) and duration (B) distributions of avalanche events during a transient period and after readaptation. 94
4.15 **Learning a Counting Task.** (A), (B) Size and duration distributions, respectively. (C) SORN_Z performance as a function of the sequence size. 96
4.16 **Learning a Sequence Task.** (A), (B) Size and duration distributions, respectively. (C) SORN_Z performance as a function of the sequence size. 98
4.17 **(Kanders et al., 2017a): Lack of co-occurrence between avalanche and edge-of-chaos criticality.** (A) Raster plots for the three dynamical regimes, from subcritical to supercritical. (B) Distributions of avalanche sizes. (C) Lyapunov spectra. 101

5.1 **Fading memory scaling for distinct dynamics in reservoirs (Bertschinger and Natschläger, 2004).** (A) Ordered, critical, and chaotic dynamics in externally driven recurrent networks. (B) Memory curve for a three-bit parity task. (C) Scaling of memory capacity. 110
5.2 **Self-organization due to plasticity improves the fading memory capacity.** 113
5.3 **Memory curves and memory capacity scaling for the SORN_L.** (A) Memory curves for a network of size $N^\text{E} = 100$, for difference alphabet sizes. (B) Equivalent curves for a $N^\text{E} = 1600$. (C) Memory capacity definition via error threshold. (D) Fading memory scaling with network size. 116

5.4 **Memory curves show an improvement in the memory capacity at medium noise levels.** 119
5.5 **Grammar learning task: training and autonomous phases.** 121
5.6 **Generation of new sentences with the $\mathbf{SORN_L}$.** (A) Incorrect and new output sentences as a function of the network size, for the FDT dictionary. (B) Incorrect and new sentences as a function of the relative number of excluded input sentences. (C) Sample autonomous output sentences for $N^{\mathrm{E}} = 200$. 122
5.7 **Errors in the generation of singular and plural sentences.** (A) Incorrect output sentences as a function of the network size. (B) Types of incorrect sentences for the SinPlu dictionary. (C) Sample autonomous output sentences for $N^{\mathrm{E}} = 200$. 125
5.8 **Statistical learning in the CHILDES dataset.** (A) Correct predictions of the subsequent character in the input corpus. (B) Character frequency in examples of spontaneous outputs. 129
5.9 **Sample spontaneous outputs for SORNs and GRUs.** . 131

B.1 **Power-law-like distributions in a log-log plot.** Comparison between randomly sampled distributions. 150

C.1 **Artificial dictionaries for the grammar learning tasks.** (A) Fox Drinks Tea dictionary (FDT). (B) Extended FDT dictionary (eFDT). (C) Singular/plural dictionary (SinPlu). . 156

Bibliography

Abbott, L. F. and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature neuroscience*, 3(11s):1178.

Agrawal, V., Cowley, A. B., Alfaori, Q., Larremore, D. B., Restrepo, J. G., and Shew, W. L. (2018). Robust entropy requires strong and balanced excitatory and inhibitory synapses. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(10):103115.

Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777.

Aswolinskiy, W. and Pipa, G. (2015). Rm-sorn: a reward-modulated self-organizing recurrent neural network. *Frontiers in computational neuroscience*, 9:36.

Atallah, B. V. and Scanziani, M. (2009). Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition. *Neuron*, 62(4):566–577.

Bak, P., Tang, C., and Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, 59(4):381.

Bak, P., Tang, C., and Wiesenfeld, K. (1988). Self-organized criticality. *Physical review A*, 38(1):364.

Bédard, C. and Destexhe, A. (2009). Macroscopic models of local field potentials and the apparent 1/f noise in brain activity. *Biophysical journal*, 96(7):2589–2603.

Bedard, C., Kroeger, H., and Destexhe, A. (2006). Does the 1/f frequency scaling of brain signals reflect self-organized critical states? *Physical review letters*, 97(11):118102.

Beggs, J. M. and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *The Journal of neuroscience*, 23(35):11167–11177.

Beggs, J. M. and Timme, N. (2012). Being critical of criticality in the brain. *Frontiers in physiology*, 3.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Bertschinger, N. and Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural computation*, 16(7):1413–1436.

Bi, G.-q. and Poo, M.-m. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of neuroscience*, 18(24):10464–10472.

Bliss, T. V. and Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of physiology*, 232(2):331–356.

Boedecker, J., Obst, O., Lizier, J. T., Mayer, N. M., and Asada, M. (2012). Information processing in echo state networks at the edge of chaos. *Theory in Biosciences*, 131(3):205–213.

Bonachela, J. A., De Franciscis, S., Torres, J. J., and Munoz, M. A. (2010). Self-organization without conservation: are neuronal avalanches generically critical? *Journal of Statistical Mechanics: Theory and Experiment*, 2010(02):P02015.

Bonachela, J. A. and Munoz, M. A. (2009). Self-organization without conservation: true or just apparent scale-invariance? *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):P09009.

Bornholdt, S. and Rohlf, T. (2000). Topological evolution of dynamical networks: Global criticality from local dynamics. *Physical Review Letters*, 84(26):6114.

Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv preprint arXiv:1206.6392*.

Bourne, J. N. and Harris, K. M. (2011). Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1 dendrites during LTP. *Hippocampus*, 21(4):354–373.

Brochini, L., de Andrade Costa, A., Abadi, M., Roque, A. C., Stolfi, J., and Kinouchi, O. (2016). Phase transitions and self-organized criticality in networks of stochastic spiking neurons. *Scientific reports*, 6.

Buonomano, D. V. and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10(2):113.

Bush, K. A. and Anderson, C. W. (2006). Exploiting iso-error pathways in the n, k-plane to improve echo state network performance. In *Neural Information Processing Systems*. Citeseer.

Chen, W., Hobbs, J. P., Tang, A., and Beggs, J. M. (2010). A few strong connections: optimizing information retention in neuronal avalanches. *BMC neuroscience*, 11(1):3.

Chialvo, D. R. (2007). The brain near the edge. In *AIP Conference Proceedings*, volume 887, pages 1–12. AIP.

Chialvo, D. R. (2010). Emergent complex neural dynamics. *Nature physics*, 6(10):744.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Churchland, M. M., Byron, M. Y., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., et al. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3):369.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Clawson, W. P., Wright, N. C., Wessel, R., and Shew, W. L. (2017). Adaptation towards scale-free dynamics improves cortical stimulus discrimination at the cost of reduced detection. *PLoS computational biology*, 13(5):e1005574.

Collins, J., Sohl-Dickstein, J., and Sussillo, D. (2016). Capacity and trainability in recurrent neural networks. *arXiv preprint arXiv:1611.09913*.

Crommelinck, M., Feltz, B., and Goujon, P. (2006). *Self-organization and emergence in life sciences*. Springer.

Dahmen, D., Diesmann, M., and Helias, M. (2016). Distributions of covariances as a window into the operational regime of neuronal networks. *arXiv preprint arXiv:1605.04153*.

Dahmen, D., Grün, S., Diesmann, M., and Helias, M. (2017). Two types of criticality in the brain. *arXiv preprint arXiv:1711.10930*.

Dan, Y., Atick, J. J., and Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of Neuroscience*, 16(10):3351–3362.

Dayan, P. and Abbott, L. F. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems.

de Andrade Costa, A., Copelli, M., and Kinouchi, O. (2015). Can dynamical synapses produce true self-organized criticality? *Journal of Statistical Mechanics: Theory and Experiment*, 2015(6):P06004.

de Arcangelis, L., Perrone-Capano, C., and Herrmann, H. J. (2006). Self-organized criticality model for brain plasticity. *Physical review letters*, 96(2):028107.

De Paola, V., Holtmaat, A., Knott, G., Song, S., Wilbrecht, L., Caroni, P., and Svoboda, K. (2006). Cell type-specific structural plasticity of axonal branches and boutons in the adult neocortex. *Neuron*, 49(6):861–875.

Dehghani, N., Hatsopoulos, N. G., Haga, Z. D., Parker, R., Greger, B., Halgren, E., Cash, S. S., and Destexhe, A. (2012). Avalanche analysis from multielectrode ensemble recordings in cat, monkey, and human cerebral cortex during wakefulness and sleep. *Frontiers in physiology*, 3:302.

Del Papa, B., Priesemann, V., and Triesch, J. (2017). Criticality meets learning: Criticality signatures in a self-organizing recurrent neural network. *PloS one*, 12(5):e0178683.

Del Papa, B., Priesemann, V., and Triesch, J. (2019). Fading memory, plasticity, and criticality in recurrent networks, in print.

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature neuroscience*, 2(6):515.

Douglas, R. J. and Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27:419–451.

Drossel, B. and Schwabl, F. (1992). Self-organized criticality in a forest-fire model. *Physica A: Statistical Mechanics and its Applications*, 191(1):47–50.

Duarte, R., Series, P., and Morrison, A. (2014). Self-organized artificial grammar learning in spiking neural networks. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 427–432.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Eser, J., Zheng, P., and Triesch, J. (2014). Nonlinear dynamics analysis of a self-organizing recurrent neural network: chaos waning. *PloS one*, 9(1):e86962.

Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292.

Fields, R. D. (2009). *The other brain: From dementia to schizophrenia, how new discoveries about the brain are revolutionizing medicine and science.* Simon and Schuster.

Font-Clos, F., Pruessner, G., Moloney, N. R., and Deluca, A. (2015). The perils of thresholding. *New Journal of Physics*, 17(4):043066.

Frémaux, N. and Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in neural circuits*, 9:85.

Frette, V., Christensen, K., Malthe-Sørenssen, A., Feder, J., Jøssang, T., and Meakin, P. (1996). Avalanche dynamics in a pile of rice. *Nature*, 379(6560):49.

Friedman, E. J. and Landsberg, A. S. (2013). Hierarchical networks, power laws, and neuronal avalanches. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(1):013135.

Friedman, N., Ito, S., Brinkman, B. A., Shimono, M., DeVille, R. L., Dahmen, K. A., Beggs, J. M., and Butler, T. C. (2012). Universal critical dynamics in high resolution neuronal avalanche data. *Physical review letters*, 108(20):208102.

Frigg, R. (2003). Self-organised criticality - what it is and what it isn't. *Studies in History and Philosophy of Science Part A*, 34(3):613–632.

Gautam, S. H., Hoang, T. T., McClanahan, K., Grady, S. K., and Shew, W. L. (2015). Maximizing sensory dynamic range by tuning the cortical state to criticality. *PLoS Comput Biol*, 11(12):e1004576.

Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B., and Massey, J. T. (1989). Mental rotation of the neuronal population vector. *Science*, 243(4888):234–236.

Gerstner, W., Kempter, R., van Hemmen, J. L., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383(6595):76.

Gireesh, E. D. and Plenz, D. (2008). Neuronal avalanches organize as nested theta-and beta/gamma-oscillations during development of cortical layer 2/3. *Proceedings of the National Academy of Sciences*, 105(21):7576–7581.

Goldstein, M. L., Morris, S. A., and Yen, G. G. (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2):255–258.

Golinkoff, R. M., Can, D. D., Soderstrom, M., and Hirsh-Pasek, K. (2015). (baby) talk to me: the social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, 24(5):339–344.

Gollo, L. L. (2017). Coexistence of critical sensitivity and subcritical specificity can yield optimal population coding. *Journal of The Royal Society Interface*, 14(134):20170207.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on*, pages 6645–6649.

Greenfield, E. and Lecar, H. (2001). Mutual information in a dilute, asymmetric neural network model. *Physical Review E*, 63(4):041905.

Haas, J. S., Nowotny, T., and Abarbanel, H. D. (2006). Spike-timing-dependent plasticity of inhibitory synapses in the entorhinal cortex. *Journal of neurophysiology*, 96(6):3305–3313.

Hahn, G., Petermann, T., Havenith, M. N., Yu, S., Singer, W., Plenz, D., and Nikolić, D. (2010). Neuronal avalanches in spontaneous activity in vivo. *Journal of neurophysiology*, 104(6):3312–3322.

Hahn, G., Ponce-Alvarez, A., Monier, C., Benvenuti, G., Kumar, A., Chavane, F., Deco, G., and Frégnac, Y. (2017). Spontaneous cortical activity is transiently poised close to criticality. *PLoS computational biology*, 13(5):e1005543.

Hahnloser, R. H., Kozhevnikov, A. A., and Fee, M. S. (2002). An ultra-sparse code underliesthe generation of neural sequences in a songbird. *Nature*, 419(6902):65.

Haimovici, A., Tagliazucchi, E., Balenzuela, P., and Chialvo, D. R. (2013). Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Physical review letters*, 110(17):178101.

Haken, H. (2008). Self-organization. *Scholarpedia*, 3(8):1401. Revision #139276.

Haldeman, C. and Beggs, J. M. (2005). Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Physical review letters*, 94(5):058101.

Han, F., Caporale, N., and Dan, Y. (2008). Reverberation of recent visual experience in spontaneous cortical waves. *Neuron*, 60(2):321–327.

Harris, T. E. (2002). *The theory of branching processes*. Courier Corporation.

Hartley, C., Taylor, T. J., Kiss, I. Z., Farmer, S. F., and Berthouze, L. (2014). Identification of criticality in neuronal avalanches: Ii. a theoretical and empirical investigation of the driven case. *Journal of mathematical neuroscience*, 4(1):1–42.

Hartmann, C., Lazar, A., Nessler, B., and Triesch, J. (2015). Where's the Noise? Key Features of Spontaneous Activity and Neural Variability Arise through Learning in a Deterministic Network. *PLoS computational biology*, 11(12):e1004640–e1004640.

Hartmann, C., Miner, D. C., and Triesch, J. (2016). Precise synaptic efficacy alignment suggests potentiation dominated learning. *Frontiers in neural circuits*, 9:90.

Hebb, D. (1949). The organization of behavior. *John wiley & sons*.

Held, G. A., Solina, D., Solina, H., Keane, D., Haag, W., Horn, P., and Grinstein, G. (1990). Experimental study of critical-mass fluctuations in an evolving sandpile. *Physical Review Letters*, 65(9):1120.

Hellyer, P. J., Jachs, B., Clopath, C., and Leech, R. (2016). Local inhibitory plasticity tunes macroscopic brain dynamics and allows the emergence of functional brain networks. *NeuroImage*, 124:85–95.

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, 3:31.

Hesse, J. and Gross, T. (2014). Self-organized criticality as a fundamental property of neural systems. *Frontiers in systems neuroscience*, 8.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoffmann, H. and Payton, D. W. (2018). Optimization by self-organized criticality. *Scientific reports*, 8(1):2358.

Holtmaat, A. and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, 10(9):647.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501.

Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, 9:133.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572.

Jaeger, H. (2002). *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach*, volume 5. GMD-Forschungszentrum Informationstechnik Bonn.

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.

Johansen-Berg, H. (2007). Structural plasticity: rewiring the brain. *Current Biology*, 17(4):R141–R144.

Jost, J. and Kolwankar, K. M. (2009). Evolution of network structure by temporal learning. *Physica A: Statistical Mechanics and its Applications*, 388(9):1959–1966.

Kadanoff, L. P. (1990). Scaling and universality in statistical physics. *Physica A: Statistical Mechanics and its Applications*, 163(1):1–14.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., of Biochemistry, D., Jessell, M. B. T., Siegelbaum, S., and Hudspeth, A. (2000). *Principles of neural science*, volume 4. McGraw-hill New York.

Kanders, K., Lorimer, T., and Stoop, R. (2017a). Avalanche and edge-of-chaos criticality do not necessarily co-occur in neural networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4):047408.

Kanders, K., Lorimer, T., Uwate, Y., Steeb, W.-H., and Stoop, R. (2017b). Robust transformations of firing patterns for neural networks. *arXiv preprint arXiv:1708.04168*.

Kanders, K. and Stoop, R. (2016). Neural avalanches at the edge-of-chaos? In *Proceedings of the 2016 International Symposium on Nonlinear Theory and its Applications (NOLTA2016)*, pages 493–496.

Karpathy, A. (2015). The unreasonable effectiveness of recurrent neural networks.

Kinouchi, O. and Copelli, M. (2006). Optimal dynamical range of excitable networks at criticality. *Nature physics*, 2(5):348–351.

Kitzbichler, M. G., Smith, M. L., Christensen, S. R., and Bullmore, E. (2009). Broadband criticality of human brain network synchronization. *PLoS computational biology*, 5(3):e1000314.

Krieg, D. and Triesch, J. (2014). A unifying theory of synaptic long-term plasticity based on a sparse distribution of synaptic strength. *Frontiers in synaptic neuroscience*, 6:3.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831.

Kuntz, M. C. and Sethna, J. P. (2000). Noise in disordered systems: The power spectrum and dynamic exponents in avalanche models. *Physical Review B*, 62(17):11699.

Lazar, A., Pipa, G., and Triesch, J. (2007). Fading memory and time series prediction in recurrent networks with different forms of plasticity. *Neural Networks*, 20(3):312–322.

Lazar, A., Pipa, G., and Triesch, J. (2009). Sorn: a self-organizing recurrent neural network. *Frontiers in computational neuroscience*, 3.

Lazar, A., Pipa, G., and Triesch, J. (2011). Emerging bayesian priors in a self-organizing recurrent network. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 127–134. Springer.

Legenstein, R. and Maass, W. (2007). Edge of chaos and prediction of computational performance for neural circuit models. *Neural Networks*, 20(3):323–334.

Legenstein, R., Pecevski, D., and Maass, W. (2008). A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS computational biology*, 4(10):e1000180.

Levina, A., Herrmann, J. M., and Geisel, T. (2007). Dynamical synapses causing self-organized criticality in neural networks. *Nature physics*, 3(12):857–860.

Levina, A., Herrmann, J. M., and Geisel, T. (2009). Phase transitions towards criticality in a neural system with adaptive interactions. *Physical review letters*, 102(11):118110.

Levina, A. and Priesemann, V. (2017). Subsampling scaling: a theory about inference from partly observed systems. *arXiv preprint arXiv:1701.04277*.

Li, W., Packard, N. H., and Langton, C. G. (1990). Transition phenomena in cellular automata rule space. *Physica. D, Nonlinear phenomena*, 45(1-3):77–94.

Loewenstein, Y., Yanover, U., and Rumpel, S. (2015). Predicting the dynamics of network connectivity in the neocortex. *Journal of Neuroscience*, 35:12535–12544.

Lombardi, F., Herrmann, H., Perrone-Capano, C., Plenz, D., and De Arcangelis, L. (2012). Balance between excitation and inhibition controls the temporal organization of neuronal avalanches. *Physical review letters*, 108(22):228703.

Lu, E. T. and Hamilton, R. J. (1991). Avalanches and the distribution of solar flares. *The astrophysical journal*, 380:L89–L92.

Lukoševičius, M. (2012). A practical guide to applying echo state networks. In *Neural networks: Tricks of the trade*, pages 659–686. Springer.

Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Löwel, S. and Singer, W. (1992). Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, 255(5041):209–212.

Maass, W. and Markram, H. (2004). On the computational power of circuits of spiking neurons. *Journal of computer and system sciences*, 69(4):593–616.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk: The database, vol. 2.

Magnasco, M. O., Piro, O., and Cecchi, G. A. (2009). Self-tuned critical anti-hebbian networks. *Physical review letters*, 102(25):258102.

Manukyan, L., Montandon, S. A., Fofonjka, A., Smirnov, S., and Milinkovitch, M. C. (2017). A living mesoscopic cellular automaton made of skin scales. *Nature*, 544(7649):173.

Marković, D. and Gros, C. (2014). Power laws and self-organized criticality in theory and nature. *Physics Reports*, 536(2):41–74.

Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, 7(2):153.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297):213–215.

Martin, C. H. and Mahoney, M. W. (2018). Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*.

Massobrio, P., de Arcangelis, L., Pasquale, V., Jensen, H. J., and Plenz, D. (2015). Criticality as a signature of healthy neural systems. *Frontiers in systems neuroscience*, 9.

Mazzoni, A., Broccard, F. D., Garcia-Perez, E., Bonifazi, P., Ruaro, M. E., and Torre, V. (2007). On the dynamics of the spontaneous activity in neuronal networks. *PloS one*, 2(5):e439.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mediano, P. A. and Shanahan, M. (2017). Balanced information storage and transfer in modular spiking neural networks. *arXiv preprint arXiv:1708.04392*.

Mehta, P. and Schwab, D. J. (2014). An exact mapping between the variational renormalization group and deep learning. *arXiv preprint arXiv:1410.3831*.

Meinhardt, H. (2008). Models of biological pattern formation: from elementary steps to the organization of embryonic axes. *Current topics in developmental biology*, 81:1–63.

Meisel, C. and Gross, T. (2009). Adaptive self-organization in a realistic neural network model. *Physical Review E*, 80(6):061917.

Meisel, C., Storch, A., Hallmeyer-Elgner, S., Bullmore, E., and Gross, T. (2012). Failure of adaptive self-organized criticality during epileptic seizure attacks. *PLoS Comput. Biol*, 8(1):e1002312.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Millman, D., Mihalas, S., Kirkwood, A., and Niebur, E. (2010). Self-organized criticality occurs in non-conservative neuronal networks during 'up' states. *Nature physics*, 6(10):801.

Milojević, S. (2010). Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology*, 61(12):2417–2425.

Miner, D. and Triesch, J. (2016). Plasticity-driven self-organization under topological constraints accounts for non-random features of cortical synaptic wiring. *PLoS Comput Biol*, 12(2):e1004759.

Minski, M. L. and Papert, S. A. (1969). Perceptrons: an introduction to computational geometry. *MA: MIT Press, Cambridge.*

Moretti, P. and Muñoz, M. A. (2013). Griffiths phases and the stretching of criticality in brain networks. *Nature communications*, 4.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100.

Neto, J. P., de Aguiar, M. A., Brum, J. A., and Bornholdt, S. (2017). Inhibition as a determinant of activity and criticality in dynamical networks. *arXiv preprint arXiv:1712.08816*.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.

Nonnenmacher, M., Behrens, C., Berens, P., Bethge, M., and Macke, J. H. (2017). Signatures of criticality arise from random subsampling in simple population models. *PLoS Computational Biology*, 13(10):e1005718.

Okun, M. and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature neuroscience*, 11(5):535.

Olami, Z., Feder, H. J. S., and Christensen, K. (1992). Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Physical Review Letters*, 68(8):1244.

Olazaran, M. (1996). A sociological study of the official history of the perceptrons controversy. *Social Studies of Science*, 26(3):611–659.

Otter, R. (1949). The multiplicative process. *The Annals of Mathematical Statistics*, pages 206–224.

Paquot, Y., Duport, F., Smerieri, A., Dambre, J., Schrauwen, B., Haelterman, M., and Massar, S. (2012). Optoelectronic reservoir computing. *Scientific reports*, 2:287.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318.

Pasquale, V., Massobrio, P., Bologna, L., Chiappalone, M., and Martinoia, S. (2008). Self-organization and neuronal avalanches in networks of dissociated cortical neurons. *Neuroscience*, 153(4):1354–1369.

Perin, R., Berger, T. K., and Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. *Proceedings of the National Academy of Sciences*, page 201016051.

Petermann, T., Thiagarajan, T. C., Lebedev, M. A., Nicolelis, M. A., Chialvo, D. R., and Plenz, D. (2009). Spontaneous cortical activity in awake monkeys composed of neuronal avalanches. *Proceedings of the National Academy of Sciences*, 106(37):15921–15926.

Plenz, D. (2013). Viewpoint: The critical brain. *Physics*, 6:47.

Poil, S.-S., Hardstone, R., Mansvelder, H. D., and Linkenkaer-Hansen, K. (2012). Critical-state dynamics of avalanches and oscillations jointly emerge from balanced excitation/inhibition in neuronal networks. *The Journal of Neuroscience*, 32(29):9817–9823.

Ponce-Alvarez, A., Jouary, A., Privat, M., Deco, G., and Sumbre, G. (2018). Whole-brain neuronal activity displays crackling noise dynamics. *Neuron*.

Priesemann, V., Munk, M. H., and Wibral, M. (2009). Subsampling effects in neuronal avalanche distributions recorded in vivo. *BMC neuroscience*, 10(1):40.

Priesemann, V., Valderrama, M., Wibral, M., and Le Van Quyen, M. (2013). Neuronal avalanches differ from wakefulness to deep sleep–evidence from intracranial depth recordings in humans. *PLoS Comput Biol*, 9(3):e1002985.

Priesemann, V., Wibral, M., Valderrama, M., Pröpper, R., Le Van Quyen, M., Geisel, T., Triesch, J., Nikolić, D., and Munk, M. H. (2014). Spike avalanches in vivo suggest a driven, slightly subcritical brain state. *Frontiers in systems neuroscience*, 8.

Privman, V. (1990). Finite-size scaling theory. *Finite Size Scaling and Numerical Simulation of Statistical Systems*, 1.

Rämö, P., Kauffman, S., Kesseli, J., and Yli-Harja, O. (2007). Measures for information propagation in boolean networks. *Physica D: Nonlinear Phenomena*, 227(1):100–104.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior*, 6(6):855–863.

Reed, W. J. and Hughes, B. D. (2002). From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical Review E*, 66(6):067103.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Ribeiro, T. L., Copelli, M., Caixeta, F., Belchior, H., Chialvo, D. R., Nicolelis, M. A., and Ribeiro, S. (2010). Spike avalanches exhibit universal dynamics across the sleep-wake cycle. *PloS one*, 5(11):e14129.

Rieke, F. and Warland, D. (1999). *Spikes: exploring the neural code.* MIT press.

Ringach, D. L. (2009). Spontaneous and driven cortical activity: implications for computation. *Current opinion in neurobiology*, 19(4):439–444.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Roxin, A., Brunel, N., Hansel, D., Mongillo, G., and van Vreeswijk, C. (2011). On the distribution of firing rates in networks of cortical neurons. *Journal of Neuroscience*, 31(45):16217–16226.

Rubinov, M., Sporns, O., Thivierge, J.-P., and Breakspear, M. (2011). Neurobiologically realistic determinants of self-organized criticality in networks of spiking neurons. *PLoS Comput Biol*, 7(6):e1002038.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533.

Sabel, B. and Schneider, G. (1988). The principle of "conservation of total axonal arborizations": massive compensatory sprouting in the hamster subcortical visual system after early tectal lesions. *Experimental brain research*, 73(3):505–518.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*.

Savin, C. and Triesch, J. (2014). Emergence of task-dependent representations in working memory circuits. *Frontiers in computational neuroscience*, 8:57.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., Van Nes, E. H., Rietkerk, M., and Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461(7260):53.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Schwab, D. J., Nemenman, I., and Mehta, P. (2014). Zipf's law and criticality in multivariate data without fine-tuning. *Physical review letters*, 113(6):068102.

Scott, G., Fagerholm, E. D., Mutoh, H., Leech, R., Sharp, D. J., Shew, W. L., and Knöpfel, T. (2014). Voltage imaging of waking mouse cortex reveals emergence of critical neuronal dynamics. *The Journal of Neuroscience*, 34(50):16611–16620.

Sethna, J. P., Dahmen, K. A., and Myers, C. R. (2001). Crackling noise. *Nature*, 410(6825):242–250.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shew, W. L., Clawson, W. P., Pobst, J., Karimipanah, Y., Wright, N. C., and Wessel, R. (2015). Adaptation to sensory input tunes visual cortex to criticality. *Nature Physics*, 11(8):659–663.

Shew, W. L. and Plenz, D. (2013). The functional benefits of criticality in the cortex. *The neuroscientist*, 19(1):88–100.

Shew, W. L., Yang, H., Petermann, T., Roy, R., and Plenz, D. (2009). Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *Journal of neuroscience*, 29(49):15595–15600.

Shew, W. L., Yang, H., Yu, S., Roy, R., and Plenz, D. (2011). Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. *The Journal of neuroscience*, 31(1):55–63.

Shi, Z. and Han, M. (2007). Support vector echo-state machine for chaotic time-series prediction. *IEEE Transactions on Neural Networks*, 18(2):359–372.

Shin, C.-W. and Kim, S. (2006). Self-organized criticality and scale-free properties in emergent functional neural networks. *Physical Review E*, 74(4):045101.

Shriki, O. (2003). *Dynamic and computational models of recurrent networks in the brain*. Hebrew University of Jerusalem.

Shriki, O., Alstott, J., Carver, F., Holroyd, T., Henson, R. N., Smith, M. L., Coppola, R., Bullmore, E., and Plenz, D. (2013). Neuronal avalanches in the resting meg of the human brain. *Journal of Neuroscience*, 33(16):7079–7090.

Shriki, O. and Yellin, D. (2016). Optimal information representation and criticality in an adaptive sensory recurrent neuronal network. *PLoS Comput Biol*, 12(2):e1004698.

Skilling, Q. M., Maruyama, D., Ognjanovski, N., Aton, S. J., and Zochowski, M. (2017). Criticality, stability, competition, and consolidation of new representations in brain networks. *arXiv preprint arXiv:1702.07649*.

Song, S., Sjöström, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS biology*, 3(3):e68.

Sornette, D. (1998). Multiplicative processes and power laws. *Physical Review E*, 57(4):4811.

Sornette, D. (2009). Dragon-kings, black swans and the prediction of crises. *arXiv preprint arXiv:0907.4290*.

Srinivasa, N., Stepp, N. D., and Cruz-Albrecht, J. (2015). Criticality as a set-point for adaptive behavior in neuromorphic hardware. *Frontiers in neuroscience*, 9:449.

Steil, J. J. (2004). Backpropagation-decorrelation: online recurrent learning with o (n) complexity. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 2, pages 843–848.

Stepp, N., Plenz, D., and Srinivasa, N. (2015). Synaptic plasticity enables adaptive self-tuning critical networks. *PLoS computational biology*, 11(1).

Stewart, C. V. and Plenz, D. (2008). Homeostasis of neuronal avalanches during postnatal cortex development in vitro. *Journal of neuroscience methods*, 169(2):405–416.

Sweeney, Y., Kotaleski, J. H., and Hennig, M. H. (2014). A diffusive homeostatic signal maintains neural heterogeneity and responsiveness in cortical networks. *bioRxiv*, page 011957.

Tagliazucchi, E., Balenzuela, P., Fraiman, D., and Chialvo, D. R. (2012). Criticality in large-scale brain fmri dynamics unveiled by a novel point process analysis. *Frontiers in Physiology*, 3(15).

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., Numata, H., Nakano, D., and Hirose, A. (2018). Recent advances in physical reservoir computing: a review. *arXiv preprint arXiv:1808.04962*.

Tanaka, T., Kaneko, T., and Aoyagi, T. (2009). Recurrent infomax generates cell assemblies, neuronal avalanches, and simple cell-like selectivity. *Neural Computation*, 21(4):1038–1067.

Taylor, T. J., Hartley, C., Simon, P. L., Kiss, I. Z., and Berthouze, L. (2013). Identification of criticality in neuronal avalanches: I. a theoretical investigation of the non-driven case. *J Math Neurosci*, 3(5).

Teixeira, F. P. P. and Shanahan, M. (2015). Local and global criticality within oscillating networks of spiking neurons. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–7.

TensorFlow (2018). Text generation using a rnn with eager execution.

Tetzlaff, C., Okujeni, S., Egert, U., Wörgötter, F., and Butz, M. (2010). Self-organized criticality in developing neuronal networks. *PLoS Comput Biol*, 6(12):e1001013.

Timme, N. M., Marshall, N. J., Bennett, N., Ripp, M., Lautzenhiser, E., and Beggs, J. M. (2016). Criticality maximizes complexity in neural tissue. *Frontiers in physiology*, 7:425.

Touboul, J. and Destexhe, A. (2010). Can power-law scaling and neuronal avalanches arise from stochastic dynamics. *PloS one*, 5(2):e8982.

Turrigiano, G. (2011). Too many cooks? intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annual review of neuroscience*, 34:89–103.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., and Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391(6670):892.

Uhlig, M., Levina, A., Geisel, T., and Herrmann, J. M. (2013). Critical dynamics in associative memory networks. *Frontiers in computational neuroscience*, 7.

Van Vreeswijk, C. and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726.

Vogels, T. P., Froemke, R. C., Doyon, N., Gilson, M., Haas, J. S., Liu, R., Maffei, A., Miller, P., Wierenga, C., Woodin, M. A., et al. (2013). Inhibitory synaptic plasticity: spike timing-dependence and putative network function. *Frontiers in neural circuits*, 7:119.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, 334(6062):1569–1573.

Watt, A. J. and Desai, N. S. (2010). Homeostatic plasticity and stdp: keeping a neuron's cool in a fluctuating world. *Frontiers in synaptic neuroscience*, 2.

Watt, A. J., van Rossum, M. C., MacLeod, K. M., Nelson, S. B., and Turrigiano, G. G. (2000). Activity coregulates quantal ampa and nmda currents at neocortical synapses. *Neuron*, 26(3):659–670.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356.

Wibral, M., Finn, C., Wollstadt, P., Lizier, J., and Priesemann, V. (2017). Quantifying information modification in developing neural networks via partial information decomposition. *Entropy*, 19(9):494.

Wibral, M., Lizier, J. T., and Priesemann, V. (2015). Bits from brains for biologically inspired computing. *Frontiers in Robotics and AI*, 2:5.

Williams-García, R. V., Moore, M., Beggs, J. M., and Ortiz, G. (2014). Quasicritical brain dynamics on a nonequilibrium widom line. *Physical Review E*, 90(6):062714.

Willits, J., Jones, M. N., and Landy, D. (2016). Learning that numbers are the same, while learning that they are different. In *CogSci*.

Willits, J., Seidenberg, M., and Saffran, J. (2009). Verbs are looking good in language acquisition. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 2570–2575.

Wilting, J., Dehning, J., Neto, J. P., Rudelt, L., Wibral, M., Zierenberg, J., and Priesemann, V. (2018). Dynamic adaptive computation: Tuning network states to task requirements. *arXiv preprint arXiv:1809.07550*.

Wilting, J. and Priesemann, V. (2016). Branching into the unknown: Inferring collective dynamical states from subsampled systems. *arXiv preprint arXiv:1608.07035*.

Xu, D., Lan, J., and Principe, J. C. (2005). Direct adaptive control: an echo state network and genetic algorithm approach. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 3, pages 1483–1486. IEEE.

Yang, H., Shew, W. L., Roy, R., and Plenz, D. (2012). Maximal variability of phase synchrony in cortical networks with neuronal avalanches. *Journal of neuroscience*, 32(3):1061–1072.

Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J., and Kasai, H. (2008). Principles of long-term dynamics of dendritic spines. *Journal of Neuroscience*, 28(50):13592–13608.

Yger, P. and Harris, K. D. (2013). The convallis rule for unsupervised learning in cortical networks. *PLoS computational biology*, 9(10):e1003272.

Yizhaq, H., Balmforth, N. J., and Provenzale, A. (2004). Blown by wind: nonlinear dynamics of aeolian sand ripples. *Physica D: Nonlinear Phenomena*, 195(3-4):207–228.

Zhang, W. and Linden, D. J. (2003). The other side of the engram: experience-driven changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience*, 4(11):885.

Zheng, P., Dimitrakakis, C., and Triesch, J. (2013). Network self-organization explains the statistics and dynamics of synaptic connection strengths in cortex. *PLoS computational biology*, 9(1):e1002848.

Zheng, P. and Triesch, J. (2014). Robust development of synfire chains from multiple plasticity mechanisms. *Frontiers in Computational Neuroscience*, 8:66.

Zierenberg, J., Wilting, J., and Priesemann, V. (2018). Homeostatic plasticity and external input shape neural network dynamics. *Physical Review X*, 8(3):031018.

Bruno Del Papa

Curriculum Vitae

Rohrbachstraße 37
60389 Frankfurt am Main, Germany
☎ *+49 176 3710 4900*
✉ *bdelpapa@gmail.com*
in *bdelpapa*
https://github.com/delpapa
Date of birth: 17.06.1990
Place of birth: São Bernardo do Campo, Brazil

General Interests

Machine learning and artificial intelligence, data science, computer vision, natural language processing; complex systems, neural networks, self-organization, computational neuroscience, information processing, reservoir computing, information theory and inference.

Education

2014–2019 **Ph.D. in Physics / Computational Neuroscience**, *International Max-Planck Research School for Neural Circuits, Frankfurt Institute for Advanced Studies, and Goethe Universität Frankfurt am Main*, *Grade: 1.0*.
Thesis Title: From criticality to learning: a study of self-organization in recurrent neural networks
Advisors: Dr. Jochen Triesch, Dr. Viola Priesemann
Frankfurt am Main, Germany

2012–2014 **M.Sc. in Physics**, *Universidade de São Paulo*, *Grade: A*.
Thesis Title: Study of the social and economic evolution of human societies through statistical mechanics and information theory methods
Advisor: Dr. Nestor Felipe Caticha Alfonso
São Paulo, Brazil

2008–2011 **B.Sc. in Physics**, *Universidade de São Paulo*, *Grade: 7.9*.
Qualification in Basic Research
São Paulo, Brazil

Additional Course Work

2018 **Deep Learning**.
5 specialization courses on deep neural networks from deeplearning.ai
Coursera

2016 **G-NODE Advanced Scientific Programming in Python**, *University of Reading*.
Reading, UK

2016 **NENGO Summer School (Brain Camp)**, *University of Waterloo*.
The NENGO Neural Simulator and Neural Engineering Framework
Waterloo, Canada

2015 **3rd Baltic-Nordic Summer School on Neuroinformatics**, *University of Tartu*.
Multiscale computational neuroscience: Neurons, networks and systems
Tartu, Estonia

2015 **Interdisciplinary College (IK)**.
From Neuron to Person: Assembling Behavior and Cognition
Möhnesee-Günne, Germany

2009 **Computability and Computational Methods**, *Universidade de São Paulo*.
University Extension Course - Programming Topics
São Paulo, Brazil

Experience

Research and development

2019–present **AI research and neuroscience**, AI researcher.
Merck KGaA, Darmstadt, Germany
Studying and developing new neuroscience inspired approaches and algorithms for unsupervised learning and with neural networks, as part of an artificial intelligence research team.

2018–2019 **Computer vision and machine learning**, Computer Vision Engineer.
Hyundai MOBIS, Frankfurt am Main, Germany
Developed deep neural network architectures for the implementation of advanced driver-assistance systems, with focus on scene classification integration and lane detection, as part of the computer vision/machine learning research and development team (*scrum* methodology). Analyzed and preprocessed various autonomous driving datasets and deep network models and evaluated their adequacy for the project needs.

2018 **Data science for the social good: Towards a More Sustainable Future of Tourism in Tuscany**, Fellow.
Nova School of Business and Economics, Lisbon, Portugal
Together with a team of four fellows, I developed innovative data driven insights to sustainable solutions for tourism mobility, with the goal of reducing the overcrowding problem in the region of Tuscany, Italy. I proposed and applied clustering and neural network models to study tourists' spatio-temporal behavior based on temporal and geospatial data (IP Probe) from millions of visitors. The project code and results are available at *http://dssg-eu.org/tuscany/index.html*

2016–2018 **Sequence and grammar learning with recurrent neural networks**.
Frankfurt Institute for Advanced Studies, Germany
I studied the spatio-temporal learning abilities of plasticity driven recurrent networks, combining Hebbian learning and homeostatic mechanisms to improve important information processing properties such as the fading memory capacity. The model is being applied for simple grammar learning tasks, and in the future might be employed for more sophisticated language processing.

2014–2018 **Stability, self-organization and learning in recurrent neural networks**.
Frankfurt Institute for Advanced Studies, Germany
As part of my PhD project, I investigated phase transitions between dynamical states in the activity of a self-organizing recurrent neural network (SORN) and the relation between criticality, learning and memory capacity in systems evolving due to biologically inspired synaptic plasticity mechanisms.

2012–2014 **Spontaneous Symmetry Breaking and Complexity of Social Networks**.
Universidade de São Paulo, Brazil
During my Master's, I investigated phase transitions in a simulated social network in order to gain insights into how complex social structures and cognitive representations can emerge due to social interaction and information exchange among agents. The project resulted in my Master thesis.

2010–2011 **Application of the Gaussian Dispersion Model**.
Instituto de Pesquisas Energéticas e Nucleares, Brazil
As an undergraduate, I worked with simulations of a stochastic model of atmospheric dispersion and applied it to estimate the amount of I^{131} deposited on the terrain near a nuclear research reactor located in the Institute de Pesquisas Energéticas e Nucleares, São Paulo, Brazil.

Teaching

2015 **Teaching Assistant**, *Frankfurt Institute for Advanced Studies*.
Course: Reinforcement Learning
Frankfurt am Main, Germany

2012-2013 **Teaching Assistant**, *Universidade de São Paulo*.
2013 - Course: Physics II
2012 - Course: Physics I
São Paulo, Brazil

2011 **Undergraduate Teaching Assistant**, *Universidade de São Paulo*.
Course: Quantum Mechanics I
São Paulo, Brazil

Scholarships and Awards

2018 *Data Science for the Social Good Fellowship*: University of Chicago

2014–2015 *PhD Fellowship*: International Max Planck Research School for Neural Circuits

2014 *PhD Scholarship (declined)*: Science without Borders, issued by the Brazilian Federal Government

2012–2014 *Master's Scholarship*: State of São Paulo Research Foundation

2012–2013 *Teaching Assistant Scholarship*: University São Paulo Teaching Improvement Program

2012 *Master's Scholarship*: Brazilian National Counsel of Technological and Scientific Development

2011 *Teaching Assistant Scholarship*: University São Paulo Undergraduate Teaching Incentive Program

2010–2011 *Undergraduate Scholarship*: Brazilian National Counsel of Technological and Scientific Development

Publications

Journals and books

2019 **Del Papa, B.**, Priesemann, V. and Triesch, J.: *Fading memory, plasticity, and criticality in neural networks.* The Functional Role of Critical Dynamics in Neural Systems. Springer (in print).

2017 **Del Papa, B.**, Priesemann, V. and Triesch, J.: *Criticality meets learning: Criticality signatures in a self-organizing recurrent neural network.* PloS ONE 12.5 (2017): e0178683.

Main conferences

2018 **Bernstein Conference**.
Talk: Learning with plasticity: from random sequences to grammar learning
Authors: Bruno Del Papa, Antonia Hufnagl, Florence Kleberg, Jochen Triesch
Berlin, Germany

2016 **Computational and Systems Neuroscience (Cosyne)**.
Poster: Criticality signatures in a self organizing recurrent neural network
Authors: Bruno Del Papa, Viola Priesemann, Jochen Triesch
Salt Lake City, USA

2015 **79th Annual Meeting of the DPG and DPG Spring Meeting**.
Deutsche Physikalische Gesellschaft eV
Poster: Neuronal avalanches in a self-organizing recurrent neural network
Authors: Bruno Del Papa, Viola Priesemann, Jochen Triesch
Berlin, Germany

2011 **XVII Undergraduate Research Seminar of the National Nuclear Energy Commission - Brazil**.
Talk and Poster: Application of the Gaussian Dispersion Model
Authors: Bruno Del Papa and Ana Maria Pinho Leite Gordon
São Paulo, Brazil

Languages

English	Full professional proficiency	*TOEFL iBT 111/120 - Jun 2013*
Portuguese	Native Speaker	
German	Intermediate proficiency	
Spanish	Intermediate proficiency	

Computer skills

Programming Languages	Python, intermediate knowledge of C/C++, MATLAB, and SQL
Software	Github, Jupyter Notebooks, BRIAN, python data analysis and machine learning packages (pytorch, tensorflow/keras, scikit-learn, numpy, scipy, pandas, geopandas, matplotlib, among others), linux systems, LaTeX, Microsoft office; basic knowledge of interactive data visualization python tools (plotly, bokeh)

Professional Affiliations

2014–2019	Frankfurt Institute of Advanced Studies
2014–2019	Max Planck Institute for Brain Research

Extracurricular Activities

2017	Volunteer at the Data Natives conference in Berlin
2015–2016	Student Representative at the Frankfurt Institute of Advanced Studies