Neural Network Growth, Structure and Dynamics

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Felipe Yaroslav Kalle Kossio
aus
Moskau, UdSSR

Bonn, 2022

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Raoul-Martin Memmesheimer
2. Gutachter:	Priv. Doz. Dr. Bernard Metsch
Tag der Promotion:	29.04.2022
Erscheinungsjahr:	2022

The evolutionary emergence of the nervous system provided individual organisms with a substrate for the development of complex behaviors and fast adaptations to dynamic environments. One of the main ways a nervous system is thought to adapt is by modifying its synapses. These are couplings between neurons, which can change their strength, appear and disappear, thus influencing the neural network dynamics and function. The plasticity rules governing changes in synaptic wiring seem to be relatively simple. This thesis examines how simple plasticity rules may lead to the development of nervous systems with complex dynamics, and how they may preserve memory despite a constant change in synaptic wiring. Finally, the thesis proposes a model for learning without synaptic modifications.

During development some nervous systems display characteristic bursts of activity called neuronal avalanches, which indicate operation near a critical point. We show using a computational model how simple plasticity rules can wire developing nervous systems such that their dynamics are close to a critical point. The simplicity of our model allows us to derive analytical expressions for the neuronal avalanche size and duration distributions. We also analyze how deviations from the critical dynamics may depend on neuronal properties such as spontaneous spiking rate and refractoriness.

Synapse are believed to be the main structures responsible for memory storage. Interestingly, however, they are not very stable; the average synapse lifetime is shorter than the lifetimes of some memories and behaviors. How can neural networks with such unstable couplings store information over long periods of time? We show that simple plasticity rules can maintain neural network function and memory despite complete rewiring of synaptic connections between neurons. In our model changes happen gradually, which allows plasticity to perform a type of error correction. Synaptic rewiring in our model also naturally leads to the change of neural representations (patterns of neural activity corresponding to particular behaviors or memories). Neural representations are often found to drift: they change in time apparently without affecting the corresponding behaviors or memories. Our analysis elucidates the mechanisms that preserve memories and behaviors despite representational drift and synaptic remodeling.

Surprisingly, there is evidence that learning of some behaviors does not require a modification of synaptic wiring. How is the information stored during such learning? We propose that a properly pretrained neural network may store the information in its dynamics eliminating the need for synaptic modifications. In our model, a network learns by modification of synapses a representative sample from a family of tasks, together with indexing tasks from that family. Later, a novel task from the same family can be learned by the network dynamics without synaptic modifications, only its identified low-dimensional index needs to be stored, possibly by neuron-intrinsic mechanisms.

Contents

CHAPTER 1

Introduction

The nervous system allows organisms to develop, learn and exhibit complex behaviors, and thus to better adapt to rich and dynamic environments during individual lives, instead of adaptation as a species on evolutionary timescales. A major goal of computational neuroscience is to construct models of nervous systems that explain behaviors, how behaviors are acquired through learning and development, and how they are maintained. A typical approach is to interconnect model neurons to form a neural network that can receive external sensory inputs and generate various outputs that result in behaviors. In general, neurons are assumed to be computationally relatively simple, with behaviors arising from their interactions that depend on the network wiring [1]. This includes behaviors ranging from simple ones like swimming of a jellyfish, to more advanced ones like rodent navigation, culminating with complex phenomena like human reason and consciousness that also seem to emerge from a large number of interacting simple neurons [2]. This thesis focuses on synaptic plasticity: the rules that govern wiring of neural networks. Synaptic plasticity plays a role throughout the life of an organism: During development, synaptic plasticity helps to shape neural networks, and later in life it is crucial for learning of new and modification of known behaviors [3]. It also plays an important part in keeping the dynamics of nervous systems stable, and maintaining the already known behaviors [4, 5]. This thesis will first examine how synaptic plasticity during development can drive networks toward a critical point where they exhibit complex dynamics. It will then demonstrate how plasticity can allow neural networks to maintain memories despite their complete rewiring. Remarkably, in both cases the plasticity rules required are quite simple. Finally, the thesis will present a model for learning without any modification of network wiring, arguably the simplest form of synaptic plasticity.

1.1 Neurons

The nervous systems of most organisms consist of numerous cell types such as neurons, astrocytes, oligodendrocytes, and many others. The non-neuronal cells, glia, account for around half of the cells in the human brain [6]. Recently some types of glia were shown to actively participate in sensory processing and behavior [7]. However, neurons are still believed to be the main actors governing behavior. Starting from sensory neurons that gather information about the environment, and ending with neurons that innervate effector tissues, neurons play the major role in the entire computation that results in behavior. This is likely due to the unique features that most neurons possess: They have an active electrical membrane that allows for fast and reliable signal transmission along long

processes (axons and dendrites). The active membrane also provides, perhaps, the most important property of neurons - a fast non-linearity: When an electrical potential across the membrane exceeds a certain threshold, for example due to the inputs from the other neurons, it spikes - rises quickly and then drops. The spikes are fast and highly stereotypic, capable of propagating along axons without decay, and of triggering synaptic transmission and thus providing inputs to other neurons. Most behaviors require non-linear computations, while a network of linear units can only perform linear transformations independent of its wiring complexity [8]. It can even be argued that a model is a good representation of a biological neural network if the biological neural activity is a linear transformation of the model activity [9]. Non-linearity allows for non-trivial computations; more than that, it allows a network to simulate any Turing machine: McCulloch and Pitts [10] showed how to construct logic gates by connecting simple model neurons that linearly sum inputs, and then apply a step non-linearity (representing linear signal integration [11] and an evoked spike). Later Siegelmann and Sontag [12] showed that neurons with sigmoidal non-linearity can be connected into a network that is a universal Turing machine. Parallel work showed that neural networks can approximate any bounded function [13] or, for a finite time, any dynamical system [14]. This computational power of networks of simple neurons provides a justification for using them: Networks made of only simple neurons may perform the same computations as networks containing more complex neurons (for example with non-linear input integration [15]) or other cell types. However, this justification should be taken with caution: Very simple systems, like Wolfram's Rule 110 cellular automata, were shown to be capable of simulating any Turing machine [16], they can therefore in principle model elaborate nervous systems, but the explanatory usefulness of such models is doubtful. In this thesis we use relatively simple, but still biologically plausible neuron models, with interesting dynamics emerging from network wiring.

1.2 Synapses

The modification of synaptic wiring, synaptic plasticity, is assumed to be the primary way networks modify their dynamics [3, 17]. Finding a connectivity that will result in the desired functioning of a model neural network is a difficult task that currently requires large amounts of data and processing power. Remarkably, biological neural networks seem to acquire correct connectivity by utilizing simple synaptic plasticity rules. Adding, removing or modifying synapses is not the only way nervous systems adapt to the environment, the adaptations happen on all levels of the structural hierarchy: New neurons (and glia) may be added to the network [18]. Properties of individual neurons may also change [19], these may include excitability, refractoriness and morphology. On a subcellular level are modifications of synapses. These are, in most cases, unidirectional chemical couplings: the spiking activity of a pre-synaptic neuron can directly influence the activity of a post-synaptic neuron. They can provide a large range of coupling strengths (also called weights) while being small structures that can be modified relatively fast. Synapses are believed to be the major elements for the adaptation of nervous systems [3, 17]. Experiments demonstrate the direct roles of synapses in adaptation: During learning, a particular set of synapses is modified and new synapses are generated. Ablation of these new and modified synapses can cause the forgetting of the learned behavior [20]. However, there are some experiments contradicting these results [21], they will be discussed later.

Synaptic plasticity is constrained by the biology of the nervous system. A synapse has access to the information that is local in space and time. The change in synaptic weight was shown to depend on the state of the pre- and post-synaptic neurons, as well as the short history of these states [3, 17].

Synapses may of course receive information from distant sources, for example via neuromodulation [17, 22], but this information channel is quite narrow, and relatively slow and unspecific. Some glial cells like astrocytes could provide an additional channel for non-local communication between synapses [23]. The synaptic plasticity rules suggested by experiments are relatively simple [3, 17, 24]. Remarkably, synaptic plasticity in invertebrates, like the sea slug *Aplysia,* that have a relatively simple nervous system and set of behaviors, seems to be conserved also in vertebrates that have much more complex brains and behaviors [25]. Although, vertebrates may also rely on additionally acquired types of plasticity. Interestingly, the underlying biological machinery supporting synaptic plasticity is complex, starting from signaling, to DNA transcription, to protein transport, and pre- and post-synapse modifications [26]. Why does such complex intracellular mechanism result in relatively simple plasticity rules between neurons? Such modularity with relatively simple interactions between modules (in our case the neurons) and complex interactions within them may be a natural property of evolved systems [27]. However, evolutionary optimization may also lead to the emergence of complex and indirect interactions between modules [28]. If that is the case for the neurons, our understanding of plasticity rules may be limited, and they may also depend on intricate interactions [29] and complex synaptic states and history [30].

It is well known, that simple rules may result in the emergence of complex systems [31]. We focus on simple plasticity rules and show that these are enough for the emergence of complex structures and dynamics during development, for learning and function maintenance. We will be mainly concerned with activity dependent plasticity: Changes in synaptic coupling strengths, that are a consequence of the neural activity. We will also utilize homeostatic plasticity: Changes of synaptic wiring that preserve the regime in which a nervous system can (optimally) function [4, 32].

1.3 Hebbian and homeostatic plasticity

Already in the early 20th century it was suggested that electrical currents reinforce paths that they take through the nervous system, although the exact mechanisms were not understood [33]. Later, synapses were identified as the major structures controlling the paths of signal propagation. It was proposed that if a synapse mediated the propagation, which is indicated first by the activity of the pre-synaptic neuron and then the activity of the paired post-synaptic neuron within a short time interval, the synaptic strength would be reinforced, facilitating further signal propagation. This is called Hebbian plasticity, after Hebb who popularized this idea. Initially, it was unclear what time interval between the pre- and post-synaptic activities will induce a change in the synaptic strength. Experiments were performed to determine the rules underlying this time dependence, the discovered plasticity was termed spike timing-dependent plasticity (STDP) [34, 35]. It was observed that when a pre-synaptic neuron spikes shortly before the post-synaptic partner, the synaptic strength increases; if the order of the spikes is reversed, the synaptic strength decreases. The exact amount of change also depends on the inter-spike interval. Since then many forms of STDP have been described [35], including ones where spike order was not important. The form of STDP has a major influence on what types of wiring patterns can emerge in the networks [36, 37]. Despite the diversity of forms, STDP rules still seem rather simple with dependence only on activity states of the pre- and post-synaptic neurons, and history effects on the order of 100 ms.

Hebbian plasticity is unstable: If a synapse is strengthened it will be more likely to mediate signal propagation in the future and thus more likely to be strengthened again, leading to a “runaway” increase

in strength. Some form of homeostatic plasticity was hypothesized to counteract this positive feedback loop and stabilize the synaptic dynamics. A prominent example is Oja's rule [38] that introduces a weight decay to stabilize the Hebbian plasticity. Interestingly, already this simple rule can perform unsupervised learning equivalent to principle component analysis. Biologically, such compensating weight decay could happen during sleep, when synaptic weights are scaled down [39]. Controlling "runaway" increase in strength is just one example of homeostatic plasticity that keeps dynamics stable. Other types of homeostatic plasticity mediate different biological processes, for example, normalization of total input synaptic strength of a neuron or control of spiking rates [4, 32].

1.4 Neuronal avalanches and critical phenomena

Can simple plasticity rules lead to the development of a neural network with non-trivial interactions and complex dynamics? A typical example of complexity is the emergence of interactions on all scales in systems with only short-range interactions [40, 41]. This type of dynamics is usually associated with a system at a critical point. In this state, a system can respond to a small perturbation on any scale up to the scale of the entire system with non-negligible probability. These responses are usually referred to as avalanches. In a critical state, the sizes of avalanches have a distribution with a power law tail. Events resembling such avalanches were first observed with multielectrode recordings of neural activity in cortical tissue slices [42], and later in developing retinas [43] and growing cultures of neurons [44]. The first cortical recordings measured local field potentials (LFP): transient electric potentials due to activities of nearby neurons [45]. The LFP came in bursts, termed neuronal avalanches, separated by quiet periods. Avalanche size was measured as a number of times the LFP at individual electrodes exceeded a certain threshold. Both neuronal avalanches size and avalanche duration distributions were reported to have power law tails, indicating a critical point. The critical point separates two phases: the phase where the neural activity quickly dies down, and another where activity would "runaway". The spiking activity of neurons, which is a major contributor to LFP, is thought to be responsible for the critical dynamics, although some experiments showed possible divergence from power law distributions when avalanches are identified from spiking data [46]. In terms of spikes, an avalanche can be thought of as a critical branching process [47], where one spike in a network produces on average one spike as an offspring. If a single spike produces on average less than one spike, the dynamics are subcritical; if a single spike produces more than one spike on average, the dynamics are supercritical.

In Chapter 2 we show how networks can reach and stay close to the critical branching point during their development. This happens because the critical point is also an attractor, such that a network can self-organize into the critical state [48]. If such a network is perturbed or initialized in a non-critical state, it converges to the critical one. We show that simple homeostatic plasticity that regulates firing rates of individual neurons is enough for the network to approach the critical point. The simplicity of the system also allows us to derive analytical expressions for avalanche size and duration distributions. Interestingly, the emergent connectivity can be changed without altering the critical dynamics by shuffling neuron's postsynaptic targets while keeping the total output structure the same. This shows that there exist many realizations of wiring that lead to the same critical dynamics.

1.5 Problems with synapses

It is generally believed that synaptic connectivity patterns determine neural network functions. The wiring is usually assumed to change in order for the neural network to acquire new functionality through learning or development. After acquiring new functionality, the wiring should remain fixed in order to preserve the function. However, it was observed that there is an ongoing spontaneous remodeling of the wiring patterns even in the absence of learning [49, 50]. These connectivity changes may be activity dependent, as well as spontaneous, independent of any network activity. Surprisingly, spontaneous remodeling does not seem to affect the functions of neural networks; in particular, memories are preserved despite changing connectivity [51, 52]. Suggestions on how stable network function may be realized with unstable synapses include: Multiple synapses between two neurons providing redundancy [53] and stability of particular synapses [54].

Because of their continuous change, synaptic wiring patterns seem ill suited for information storage. Additionally, there is evidence indicating that learning or memory in some cases do not require synapses or their modifications: Purkinje cells learning in a classical conditioning paradigm was reported to happen on the cellular rather than the network level [55]. Cultures of *Aplysia* neurons retained memory despite erasure of synaptic modifications associated with learning [21]. Mice could acquire a novel task while synaptic modifications were blocked, if this task was related to a previously learned one [56, 57]. Monkeys learned reaching tasks but showed no signs of altered functional connectivity [58]. Such evidence, together with constant change in wiring and the simplicity of plasticity rules, might suggest that synaptic wiring patterns may not store information nor determine network functions [59]. Proposals for alternative long-term information storage include epigenetic mechanisms [60, 61], since they are already utilized as a long-term storage on the cellular level, and a long-living extracellular molecular structure of the perineuronal nets [62].

When a neural network receives input from sensory organs, the activity within it reflects, or represents, this input. Similarly, when the network sends commands to effector tissues, the activity should represent these commands. And when a memory is recalled, the activity of the neurons should reflect that also. A neural representation is a neural activity pattern that correlates to a certain physical reality, like presentation of stimulus, memory recall, or behavior. Surprisingly, memory representations and other representations are not static, they evolve in time [63–65]. Neurons active during certain behavior or in response to stimuli may stop being active or change their response at a later time. This can happen if memories are modified [66]. More surprisingly, it can also happen without any observable change in behavior [64, 65]. The change of neural representations also happens on time scales shorter than memory lifetime, and is another challenge for the static connectivity patterns as the determinants of a network function [52, 65].

How can a network with drifting representations reliably perform behaviors and recall memories? For example, a motor neuron that initiates muscle contraction needs to faithfully readout the information about the planned motion trajectory from the neurons representing the trajectories, but these representations drift. There are a few suggestions on how correct readouts can be performed from drifting representations [52]: First, the change in representation may be incomplete, some neurons may be stable and continue to participate in the representation [67], by relying on them a readout may correctly decode the representation. So far experiments found that a stable fraction of neurons remains in representations, but it is unclear whether longer term neural recordings would show otherwise. Second, redundant representations may exist in the neural network, the switching between them may appear as representational drift [68]. Evidence against this is the reduced behavioral response when

neurons that formed a representation at an earlier time are artificially stimulated [64]. Finally, it is also possible that readout neurons receive some error feedback to relearn the representation as it drifts [69]: The same mechanism used during novel learning could be used to relearn the drifting representation. This requires error feedback that comes from repeated behavior. However, experiments showed normal behavioral performance despite drift even if behaviors were not repeated between learning and recall [64].

1.6 Drifting neuronal assemblies

To answer how neural networks may preserve functions despite synaptic rewiring and drifting representations, we consider, in Chapter 3, simple memory representations called neuronal assemblies. Neuronal assemblies are one of the earliest models for storing associative memories. They are groups of neurons strongly coupled relative to the average network coupling [70, 71]. If some of the assembly neurons are active, the rest of the assembly will also become active due to the strong coupling. This complete assembly activation after partial stimulation models the cued recall of associative memory. Assemblies may be created during learning with the help of external input [72]. There is also evidence that already during development assemblies may be formed with very little sensory input [73]. In both cases simple plasticity rules are enough for the formation of assemblies. We demonstrate by using standard Hebbian and homeostatic plasticity rules together with assemblies as memory representations that the representational drift and network rewiring can emerge naturally without disrupting stored memories. Our models demonstrate how drifting assemblies may be reliably read out without requiring a stable group of neurons, multiple redundant states, or error feedback. They also explain the origin of the representational drift and identify spontaneous synaptic rewiring as one of its drives.

1.7 Learning without synaptic modification

Drifting assemblies that compensate for change using Hebbian and homeostatic plasticity illustrate how functions may be preserved despite constantly changing network wiring. In Chapter 4 we address how learning may happen without requiring synaptic modification, as shown by some experiments. In particular it was observed that mice failed to learn navigation tasks if modifications of one type of synapse were prevented [56, 57, 74]. However, if mice were already familiar with a similar navigation task, a new task could be learned even with blocked synaptic modifications. Inspired by this we develop a method to pretrain networks allowing synaptic modification, on samples from a family of tasks such that the network can later learn new tasks from that family without synaptic modifications. We focus on supervised dynamical learning of trajectories and dynamical systems. A network is presented with an error signal that shows how far the currently generated trajectory is from the desired one. After the error signal is turned off, the network continues to generate the required trajectory. The learning happens through the modification of the network dynamics, constant activity patterns of “context” neurons store the learned task parameters. For simple tasks this constant activity is low dimensional and may be stored in a neuron intrinsic manner for long-term memories, for example as suggested in [55].

References

[1] H. Sompolinsky, *Computational neuroscience: beyond the local circuit*, Current Opinion in Neurobiology **25** (2014) xiii.

[2] S. Herculano-Houzel, *The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost*, Proceedings of the National Academy of Sciences **109** (2012) 10661.

[3] A. Citri and R. C. Malenka, *Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms*, Neuropsychopharmacology **33** (2008) 18.

[4] G. G. Turrigiano, *Homeostatic plasticity in neuronal networks: the more things change, the more they stay the same*, Trends in Neurosciences **22** (1999) 221.

[5] J. R. Wolpaw, *Spinal cord plasticity in acquisition and maintenance of motor skills*, Acta Physiologica **189** (2007) 155.

[6] C. S. von Bartheld, J. Bahney and S. Herculano-Houzel, *The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting*, Journal of Comparative Neurology **524** (2016) 3865.

[7] P. Kofuji and A. Araque, *Astrocytes and Behavior*, Annual Review of Neuroscience **44** (2021) 49.

[8] A. M. Zador, *The basic unit of computation*, Nature Neuroscience **3** (2000) 1167.

[9] P. Bashivan, K. Kar and J. J. DiCarlo, *Neural population control via deep image synthesis*, Science **364** (2019) eaav9436.

[10] W. S. McCulloch and W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, The bulletin of mathematical biophysics **5** (1943) 115.

[11] S. Cash and R. Yuste, *Linear Summation of Excitatory Inputs by CA1 Pyramidal Neurons*, Neuron **22** (1999) 383.

[12] H. T. Siegelmann and E. D. Sontag, *On the computational power of neural nets*, Journal of computer and system sciences **50** (1995) 132.

[13] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals and Systems **2** (1989) 303.

[14] K.-i. Funahashi and Y. Nakamura, *Approximation of dynamical systems by continuous time recurrent neural networks*, Neural Networks **6** (1993) 801.

[15] B. B. Ujfalussy, J. K. Makara, T. Branco and M. Lengyel, *Dendritic nonlinearities are tuned for efficient spike-based computations in cortical circuits*, eLife **4** (2015) e10056, ed. by F. K. Skinner.

[16] M. Cook et al., *Universality in elementary cellular automata*, Complex systems **15** (2004) 1.

[17] J. C. Magee and C. Grienberger, *Synaptic Plasticity Forms and Functions*, Annual Review of Neuroscience **43** (2020) 95.

[18] E. C. Cope and E. Gould, *Adult Neurogenesis, Glia, and the Extracellular Matrix*, Cell Stem Cell **24** (2019) 690.

[19] E. Marder, L. F. Abbott, G. G. Turrigiano, Z. Liu and J. Golowasch,
Memory from the dynamics of intrinsic membrane currents,
Proceedings of the National Academy of Sciences **93** (1996) 13481.

[20] M.-m. Poo et al., *What is memory? The present state of the engram*,
BMC Biology **14** (2016) 40.

[21] S. Chen et al., *Reinstatement of long-term memory following erasure of its behavioral and synaptic expression in Aplysia*, eLife **3** (2014) e03896, ed. by M. Ramaswami.

[22] Z. Brzosko, S. B. Mierau and O. Paulsen,
Neuromodulation of Spike-Timing-Dependent Plasticity: Past, Present, and Future,
Neuron **103** (2019) 563.

[23] G. Perea, M. Navarrete and A. Araque,
Tripartite synapses: astrocytes process and control synaptic information,
Trends in Neurosciences **32** (2009) 421.

[24] B. A. Hassan and P. R. Hiesinger,
Beyond Molecular Codes: Simple Rules to Wire Complex Brains, Cell **163** (2015) 285.

[25] D. L. Glanzman, *Common Mechanisms of Synaptic Plasticity in Vertebrates and Invertebrates*,
Current Biology **20** (2010) R31.

[26] E. R. Kandel,
The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses,
Science **294** (2001) 1030.

[27] J. Clune, J.-B. Mouret and H. Lipson, *The evolutionary origins of modularity*,
Proceedings of the Royal Society B: Biological Sciences **280** (2013) 20122863.

[28] A. Thompson, "An evolved circuit, intrinsic in silicon, entwined with physics",
Evolvable Systems: From Biology to Hardware, ed. by T. Higuchi, M. Iwata and W. Liu,
Berlin, Heidelberg: Springer Berlin Heidelberg, 1997 390, ISBN: 978-3-540-69204-1.

[29] J. Ashley et al.,
Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons,
Cell **172** (2018) 262.

[30] S. Fusi and L. F. Abbott, *Limits on the memory storage capacity of bounded synapses*,
Nature Neuroscience **10** (2007) 485.

[31] S. Wolfram, *A New Kind of Science*, English, Wolfram Media, 2002, ISBN: 1579550088,
URL: `https://www.wolframscience.com`.

[32] N. Vitureira, M. Letellier and Y. Goda,
Homeostatic synaptic plasticity: from single synapses to neural circuits,
Current Opinion in Neurobiology **22** (2012) 516.

[33] G. Berlucchi and H. A. Buchtel, *Neuronal plasticity: historical roots and evolution of meaning*,
Experimental Brain Research **192** (2009) 307.

[34] Y. Dan and M.-m. Poo, *Spike Timing-Dependent Plasticity of Neural Circuits*,
Neuron **44** (2004) 23.

[35] N. Caporale and Y. Dan, *Spike Timing–Dependent Plasticity: A Hebbian Learning Rule*, Annual Review of Neuroscience **31** (2008) 25.

[36] N. Ravid Tannenbaum and Y. Burak, *Shaping Neural Circuits by High Order Synaptic Interactions*, PLoS Computational Biology **12** (2016) 1.

[37] G. K. Ocker, A. Litwin-Kumar and B. Doiron, *Self-Organization of Microcircuits in Networks of Spiking Neurons with Plastic Synapses*, PLoS Computational Biology **11** (2015) 1.

[38] E. Oja, *Simplified neuron model as a principal component analyzer*, Journal of Mathematical Biology **15** (1982) 267.

[39] G. Tononi and C. Cirelli, *Sleep function and synaptic homeostasis*, Sleep Medicine Reviews **10** (2006) 49.

[40] P. Bak and M. Paczuski, *Complexity, contingency, and criticality*, Proceedings of the National Academy of Sciences **92** (1995) 6689.

[41] P. Bak, K. Chen and M. Creutz, *Self-organized criticality in the 'Game of Life"*, Nature **342** (1989) 780.

[42] J. M. Beggs and D. Plenz, *Neuronal Avalanches in Neocortical Circuits*, Journal of Neuroscience **23** (2003) 11167.

[43] M. H. Hennig, C. Adams, D. Willshaw and E. Sernagor, *Early-Stage Waves in the Retinal Network Emerge Close to a Critical State Transition between Local and Global Functional Connectivity*, Journal of Neuroscience **29** (2009) 1077.

[44] Y. Yada et al., *Development of neural population activity toward self-organized criticality.*, Neuroscience **343** (2017) 55.

[45] G. Buzsáki, C. A. Anastassiou and C. Koch, *The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes*, Nature Reviews Neuroscience **13** (2012) 407.

[46] J. Wilting and V. Priesemann, *25 years of criticality in neuroscience — established results, open controversies, novel concepts*, Current Opinion in Neurobiology **58** (2019) 105.

[47] S. Zapperi, K. B. Lauritsen and H. E. Stanley, *Self-Organized Branching Processes: Mean-Field Theory for Avalanches*, Physical Review Letters **75** (22 1995) 4071.

[48] P. Bak, C. Tang and K. Wiesenfeld, *Self-organized criticality*, Physical Review A **38** (1 1988) 364.

[49] N. E. Ziv and N. Brenner, *Synaptic Tenacity or Lack Thereof: Spontaneous Remodeling of Synapses*, Trends in Neurosciences **41** (2018) 89.

[50] H. Kasai, N. E. Ziv, H. Okazaki, S. Yagishita and T. Toyoizumi, *Spine dynamics in the brain, mental disorders and artificial neural networks*, Nature Reviews Neuroscience **22** (2021) 407.

[51] G. Mongillo, S. Rumpel and Y. Loewenstein, *Intrinsic volatility of synaptic connections — a challenge to the synaptic trace theory of memory*,
Current Opinion in Neurobiology **46** (2017) 7.

[52] A. R. Chambers and S. Rumpel, *A stable brain from unstable components: Emerging concepts and implications for neural computation*, Neuroscience **357** (2017) 172.

[53] M. Fauth, F. Wörgötter and C. Tetzlaff, *Formation and Maintenance of Robust Long-Term Information Storage in the Presence of Synaptic Turnover*,
PLoS Computational Biology **11** (2016) 1.

[54] G. Mongillo, S. Rumpel and Y. Loewenstein,
Inhibitory connectivity defines the realm of excitatory plasticity,
Nature Neuroscience **21** (2018) 1463.

[55] F. Johansson, D.-A. Jirenhed, A. Rasmussen, R. Zucca and G. Hesslow,
Memory trace and timing mechanism localized to cerebellar Purkinje cells,
Proceedings of the National Academy of Sciences **111** (2014) 14930.

[56] D. Saucier and D. P. Cain,
Spatial learning without NMDA receptor-dependent long-term potentiation,
Nature **378** (1995) 186.

[57] M. K. Otnæss, V. H. Brun, M.-B. Moser and E. I. Moser, *Pretraining Prevents Spatial Learning Impairment after Saturation of Hippocampal Long-Term Potentiation*,
Journal of Neuroscience **19** (1999) RC49.

[58] M. G. Perich, J. A. Gallego and L. E. Miller,
A Neural Population Mechanism for Rapid Learning, Neuron **100** (2018) 964.

[59] P. C. Trettenbrein,
The Demise of the Synapse As the Locus of Memory: A Looming Paradigm Shift?,
Frontiers in Systems Neuroscience **10** (2016) 88.

[60] A. Bédécarrats, S. Chen, K. Pearce, D. Cai and D. L. Glanzman, *RNA from Trained Aplysia Can Induce an Epigenetic Engram for Long-Term Sensitization in Untrained Aplysia*,
eNeuro **5** (2018).

[61] R. Holliday, *Is there an Epigenetic Component in Long-term Memory?*,
Journal of Theoretical Biology **200** (1999) 339.

[62] R. Y. Tsien,
Very long-term memories may be stored in the pattern of holes in the perineuronal net,
Proceedings of the National Academy of Sciences **110** (2013) 12456.

[63] C. E. Schoonover, S. N. Ohashi, R. Axel and A. J. P. Fink,
Representational drift in primary olfactory cortex, Nature **594** (2021) 541.

[64] L. A. DeNardo et al.,
Temporal evolution of cortical ensembles promoting remote memory retrieval,
Nature Neuroscience **22** (2019) 460.

[65] M. E. Rule, T. O'Leary and C. D. Harvey, *Causes and consequences of representational drift*,
Current Opinion in Neurobiology **58** (2019) 141.

[66] A. Tompary and L. Davachi, *Consolidation Promotes the Emergence of Representational Overlap in the Hippocampus and Medial Prefrontal Cortex*, Neuron **96** (2017) 228.

[67] C. Clopath, T. Bonhoeffer, M. Hübener and T. Rose, *Variance and invariance of neuronal long-term representations*, Philosophical Transactions of the Royal Society B: Biological Sciences **372** (2017) 20160161.

[68] U. Rokni, A. G. Richardson, E. Bizzi and H. S. Seung, *Motor Learning with Unstable Neural Representations*, Neuron **54** (2007) 653.

[69] M. E. Rule et al., *Stable task information from an unstable neural population*, eLife **9** (2020) e51121, ed. by S. Palmer and R. L. Calabrese.

[70] A. Scott, *Neuroscience: a Mathematical Primer*, Springer New York, 2002, URL: https://doi.org/10.1007/b98897.

[71] G. Buzsáki, *Neural Syntax: Cell Assemblies, Synapsembles, and Readers*, Neuron **68** (2010) 362.

[72] A. Litwin-Kumar and B. Doiron, *Formation and maintenance of neuronal assemblies through synaptic plasticity.*, eng, Nature Communications **5** (2014) 5319.

[73] M. A. Triplett, L. Avitan and G. J. Goodhill, *Emergence of spontaneous assembly activity in developing neural networks without afferent input*, PLoS Computational Biology **14** (2018) 1.

[74] D. M. Bannerman, M. A. Good, S. P. Butcher, M. Ramsay and R. G. M. Morris, *Distinct components of spatial learning revealed by prior training and NMDA receptor blockade*, Nature **378** (1995) 182.

CHAPTER 2

Growing Critical: Self-Organized Criticality in a Developing Neural System

Experiments in various neural systems found avalanches: bursts of activity with characteristics typical for critical dynamics. A possible explanation for their occurrence is an underlying network that self-organizes into a critical state. We propose a simple spiking model for developing neural networks, showing how these may "grow into" criticality. Avalanches generated by our model correspond to clusters of widely applied Hawkes processes. We analytically derive the cluster size and duration distributions and find that they agree with those of experimentally observed neuronal avalanches.

This chapter is a reproduction with minor alterations of the article of the same title that was published in Physical Review Letters under the reference: Felipe Yaroslav Kalle Kossio, Sven Goedeke, Benjamin van den Akker, Borja Ibarz, and Raoul-Martin Memmesheimer, Phys. Rev. Lett. 121, 058301, https://doi.org/10.1103/PhysRevLett.121.058301. The supplementary material is reproduced in Appendix A.

2.1 Introduction

A hallmark of systems at criticality is the variability of their responses to small perturbations. While small responses are most likely, the probability of large, system-size effects is non-negligible. Various natural and model complex systems show similar behavior [1]. One explanation is that they drive themselves close to a critical state ("self-organized criticality" [2, 3]). The dynamics of such systems are characterized by "events" or "avalanches". Their sizes and durations follow power-law distributions, frequently with exponents $3/2$ and 2, indicating an underlying critical branching process [4–7]. Apparent critical dynamics, "neuronal avalanches", in biological neural networks were first reported in Refs. [8, 9]. It has been suggested that they foster information storage and transfer [10, 11]. Experimental studies often report power-law size and duration distributions with exponents $3/2$ and 2. They further indicate that neuronal avalanches emerge during development [12–15], suggesting that neural networks develop into a critical state.

The development of neural networks is determined by an interplay of genetic determinants and environmental influence. Of pivotal importance is neural activity [16, 17]. As a general rule, neurons with low activity level extend their neurites and form more activating connections, while highly active cells reduce these [18–20]. Thereby, neurons maintain their average activity at a particular level (homeostasis) [21–23].

Computational models for avalanches in neural systems rely on static, tuned connectivity [14, 24], on short-term synaptic plasticity [25, 26], or on long-term network changes [27–30]. Here we propose a continuous-time spiking neural network model belonging to the third class. The avalanche dynamics follow from a network growth process towards a critical state, which uses local information only [31–33]. The model is rooted in previous models for neural network development [27, 29, 34], but sufficiently simple to be analytically tractable.

2.2 Neuron model

Like biological neurons, our model neurons communicate by sending and receiving spikes in continuous time. Spiking is stochastic, according to an inhomogeneous Poisson point process with instantaneous rate $f_i(t)$ for neuron i [27, 35–38]. In isolation neurons have a low spontaneous rate f_0, e.g., due to spontaneous synaptic release or channel fluctuations [39, 40]. A spike from neuron j increases f_i by the size of the time-dependent connection strength gA_{ij}. The increment decays exponentially with time constant τ, which accounts for relaxation due to leak currents. The couplings are excitatory; this is the dominant connection type in developing neural systems [34]. Taken together, f_i's dynamics follow

$$\tau \dot{f}_i(t) = f_0 - f_i(t) + \tau g \sum_j A_{ij}(t^-) \sum_{\hat{t}_j} \delta\left(t - \hat{t}_j\right), \tag{2.1}$$

where $\hat{t}_j$ denotes the spike times of neuron j (δ is the Dirac delta distribution). For simplicity, we assume that all neurons have the same parameters. For constant couplings, the network dynamics form a multivariate Hawkes point process [37, 38, 41].

2.3 Network growth

Neurons are commonly arranged in single or stacked layers. We thus represent neurite extents by disks with radii $R_i(t)$, with centers, representing cell somas, randomly and uniformly distributed in a planar area [27, 34, 42]. Since neurons with more neurite overlap can grow more synaptic connections [43, 44], connection strengths are set proportional to the overlap areas $A_{ij}(t)$ of the disks, with proportionality constant g. We incorporate homeostatic neurite growth by evolving extents as

$$\dot{R}_i(t) = K\left(1 - \frac{1}{f_{\text{sat}}} \sum_{\hat{t}_i} \delta(t - \hat{t}_i)\right), \tag{2.2}$$

Fig. 2.1. Between spikes of neuron i, $R_i(t)$ grows linearly with rate K. At spike sending, it shrinks by a constant amount K/f_{sat}, which determines the rate f_{sat} at which growth and shrinkage equilibrate. There are no self-connections. Growth takes much longer than decay of activity, $1/K \gg \tau$ (spatial scales of the population are of order one). Furthermore, we assume $f_{\text{sat}} \gg f_0$, in agreement with

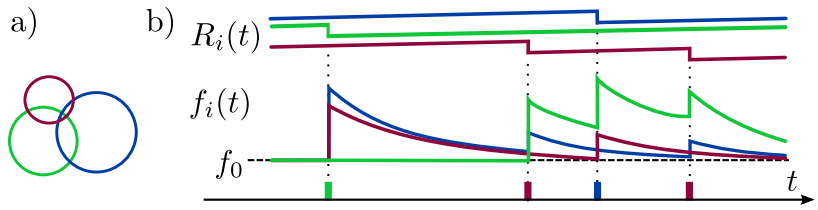

Figure 2.1: Neuron dynamics. (a) Neurons' somas and neurite extents are represented by disks with evolving radii. Coupling strengths are proportional to neurite overlap areas. (b) Neurite radii increase linearly (upper traces), own spike sendings (lower trace) result in instantaneous shrinkage. Spike arrivals increase the instantaneous firing rate by the coupling strength, it decays exponentially in between (middle trace).

experiments [15, 40], Appendix A.2. Spontaneously inactive neurons would reduce the relevant average f_0, Appendix A.4. The growth model is biologically inspired; it is a simplification of previous growth models [23, 27, 29, 34]. However, many slow homeostatic processes [21, 22, 45] with $f_{\text{sat}} \gg f_0$ will yield similar results.

The neurons are initially mostly isolated. Over time, they extend their neurites, form connections, and develop a network, Fig. 2.2. At intermediate stages, neurites and overlaps can overshoot [15, 29, 34]. When neuron i's time-averaged firing rate $\bar{f}_i$ reaches f_{sat}, its average growth stops [Eq. (2.2)]. Our networks grow into a stationary state, where $\bar{f}_i = f_{\text{sat}}$ for all i. In the following, we investigate this state.

2.4 Stationary state dynamics

We first compute the average number of spikes that a spike directly causes: Identical $\bar{f}_i$ imply identical time-averaged total overlaps $\sum_j \bar{A}_{ij} = \bar{A}_i = \bar{A}$ and input coupling strengths. Time averaging Eq. (2.1), $\bar{f}_i = f_0 + \tau g \sum_j \bar{A}_{ij} \bar{f}_j$ [here and henceforth we neglect the small fluctuations of $A_{ij}(t)$ around $\bar{A}_{ij}$], and inserting f_{sat} yields $\tau g \bar{A} = 1 - f_0/f_{\text{sat}}$. A spike of neuron j at $\hat{t}_j$ adds $g A_{ij}(\hat{t}_j^-) e^{-(t-\hat{t}_j)/\tau} \Theta(t - \hat{t}_j)$ to $f_i(t)$ [Eq. (2.1), Θ is the Heaviside function], such that the number of additionally induced spikes in neuron i is Poisson distributed with mean $\tau g A_{ij}(\hat{t}_j^-)$. Averaged over the randomness of spike generation each spike thus generates in total

$$\sigma = \tau g \sum_i \bar{A}_{ij} = \tau g \bar{A} = 1 - \frac{f_0}{f_{\text{sat}}} \tag{2.3}$$

spikes, where we used the symmetry of overlaps, $A_{ij} = A_{ji}$, $\sum_i \bar{A}_{ij} = \sum_j \bar{A}_{ij} = \bar{A}$. Equation (2.3) holds independently of network activity and neuron identity, due to the linearity of Eq. (2.1) and the homogeneity of parameters. In particular, σ equals also the time and population average number of induced spikes ($\propto f_{\text{sat}} - f_0$) per spike ($\propto f_{\text{sat}}$).

The independence of spike offspring generation from other spikes allows us to understand the dynamics as a branching process with branching parameter σ. More specifically, we have an age-dependent or Crump-Mode-Jagers branching process [46]: Individuals (spikes) generate offspring at an age-dependent rate. Neuronal avalanches are trees of offspring, started by a spontaneous spike. For their overall size only the distribution of single spike offspring matters. It is Poissonian with parameter

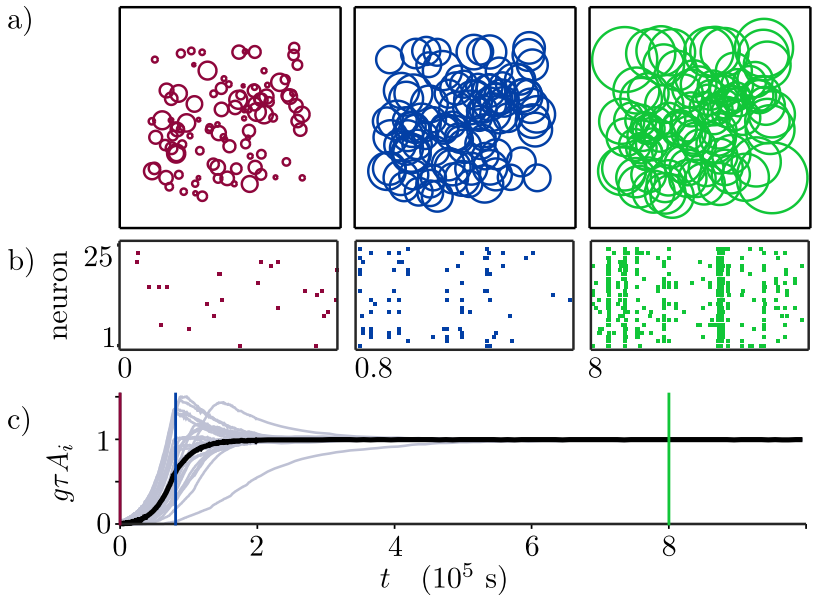

Figure 2.2: Network dynamics. (a) Extents of neurites. (b) Spikes generated by 25 sample neurons (100 s windows). (c) Scaled total overlaps of 25 sample neurons (gray) and the population average (black) as a function of time. For (a) and (b) from left to right: initial state (red), state with growth on average (blue), stationary state (green). Color coded vertical lines in (c) indicate the three different time points in (a) and (b).

σ. The avalanche sizes s therefore follow the Borel distribution [47],

$$P(s) = \frac{(s\sigma)^{s-1} e^{-s\sigma}}{s!}. \tag{2.4}$$

We apply Stirling's approximation to obtain

$$P_{\text{appr}}(s) = \frac{1}{\sqrt{2\pi\sigma}} s^{-3/2} e^{-(\sigma - \ln\sigma - 1)s}, \tag{2.5}$$

explicitly highlighting the power-law tail with exponent $3/2$ of a critical branching process for $\sigma = 1$ [4–7]. For a subcritical process ($\sigma < 1$), Eq. (2.5) is a power law with exponential cutoff around $s_c(\sigma) = (\sigma - \ln\sigma - 1)^{-1}$. It signifies subcritical dynamics [4, 5, 48], not a finite size effect [3, 5]; the size distribution is independent of neuron number. Equation (2.5) inherits the good quality of Stirling's approximation [49], with relative error about $1/(12s)$.

The heights of Crump-Mode-Jagers trees, i.e., the temporal differences T between their first and last individuals, represent the durations of the corresponding neuronal avalanches. In the following we derive their probability density $p(T)$. Because of the additivity of Poisson processes, the superposition of all neurons' spike trains can be described as an inhomogeneous Poisson process with rate $f(t) = \sum_i f_i(t)$. Summing Eq. (2.1) over i and inserting $\bar{A} = \sigma/(g\tau)$ [Eq. (2.3)] yields $\tau\dot{f}(t) = Nf_0 - f(t) + \sigma\sum_{\hat{t}} \delta(t - \hat{t})$ with the number of neurons N. $\hat{t}$ are the neurons' spike times; they occur with instantaneous rate $f(t)$. The spiking dynamics may thus be interpreted as a self-exciting Hawkes process. It is Markovian due to the exponentially decaying impact kernel [50, 51]. The spontaneous background rate Nf_0 initiates avalanches. To determine their durations, we consider an

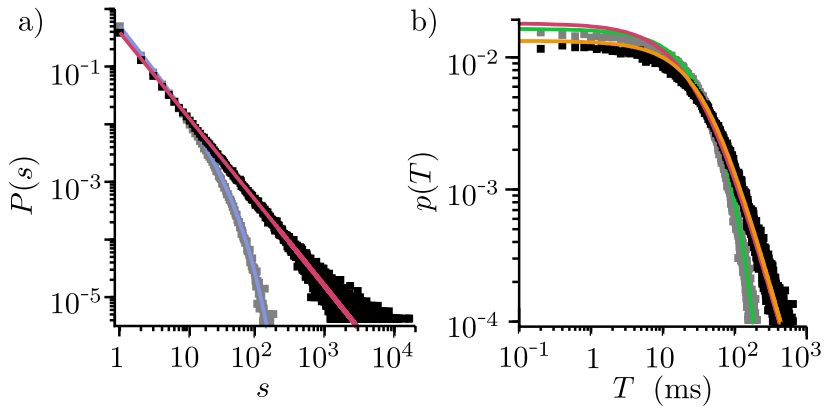

Figure 2.3: Avalanche sizes and durations. (a) Analytical size distributions Eq. (2.5) (discrete points connected) and simulation results for subcritical ($f_{\text{sat}} = 0.04\,\text{Hz}$, $\sigma = 0.75$, $t_{\text{bin}} = 30\,\text{ms}$, blue and gray) and near-critical ($f_{\text{sat}} = 2\,\text{Hz}$, $\sigma = 0.995$, $t_{\text{bin}} = 45\,\text{ms}$, red and black) states. Equation (2.4) yields visually indistinguishable analytics. (b) Analytical duration distributions Eq. (2.10) and simulation results, for subcritical (green and gray) and near-critical (orange and black) states, and closed-form approximation Eq. (2.12) (red).

analogous process with instantaneous rate $f_a(t)$ and without spontaneous spiking, which is initiated at $t = 0$ by an external spike,

$$\tau \dot{f}_a(t) = -f_a(t) + \sigma \sum_{\hat{t}_a} \delta(t - \hat{t}_a), \;\; f_a(0) = \frac{\sigma}{\tau}. \tag{2.6}$$

The duration of an avalanche is the time T of this process' last spike. The probability that it has occurred before t gives the distribution function $P(T \leq t)$ of durations. We first compute this probability conditioned on the instantaneous rate $f_a(t)$ at the end of the considered interval:

$$\begin{aligned} P(T \leq t \mid f_a(t)) &= P(\text{no spike in } (t, \infty) \mid f_a(t)) \\ &= e^{-\int_t^\infty f_a(t')dt'} = e^{-\tau f_a(t)}, \end{aligned} \tag{2.7}$$

where we use that the process behaves like a Poisson process with exponentially decaying rate, if no spike is generated. Averaging over $f_a(t)$ yields

$$\begin{aligned} P(T \leq t) &= \int_0^\infty P\left(T \leq t \mid f_a(t)\right) p\left(f_a(t)\right) df_a(t) \\ &= E\left(e^{-\tau f_a(t)}\right). \end{aligned} \tag{2.8}$$

$E(\cdot)$ denotes the expectation value over the process, Eq. (2.6). Importantly, Eq. (2.8) shows that $P(T \leq t)$ equals the Laplace transform of the random variable $f_a(t)$, evaluated at the decay time τ. This Laplace transform has recently been derived [52–54]. Inserting our parameters yields $E(e^{-\tau f_a(t)}) = e^{\sigma a(t)/\tau}$, where $a(t)$ satisfies

$$\dot{a}(t) = -a(t)/\tau + e^{\sigma a(t)/\tau} - 1, \;\; a(0) = -\tau. \tag{2.9}$$

The resulting $P(T \leq t) = e^{\sigma a(t)/\tau}\Theta(t)$ with $\Theta(0) = 1$ has the density

$$p(T) = \sigma \dot{a}(T) e^{\sigma a(T)/\tau}\Theta(T)/\tau + e^{-\sigma}\delta(T). \tag{2.10}$$

We can generalize Eq. (2.9) to Hawkes processes with different kernels using the integral equation for cluster duration distributions [55, 56].

We finally approximate $p(T)$ by closed-form expressions with a focus on its tail near criticality. For large t, $P(T \leq t)$ approaches 1, so $a(t)$ approaches 0. Generally, $\sigma a(t)/\tau$ stays between -1 and 0. Expanding $e^{\sigma a(t)/\tau}$ in Eq. (2.9) around $\sigma a(t)/\tau = 0$ to second order,

$$\dot{a}(t) \approx (\sigma - 1)a(t)/\tau + \sigma^2 a(t)^2/(2\tau^2),\ a(0) = -\tau, \tag{2.11}$$

yields closed-form approximations for $a(t)$. In particular, for nearly critical systems with $\sigma \approx 1$, the first term on the right-hand side vanishes and the solution is $a_{\text{appr}}(t) = -2\tau^2/(2\tau + t)$, leading to a probability density

$$p_{\text{appr}}(T) = 2\tau(2\tau + T)^{-2} e^{a_{\text{appr}}(T)/\tau}\Theta(T) + e^{-1}\delta(T), \tag{2.12}$$

which approaches for large T a power law with critical exponent 2. For large t the error in the expansion Eq. (2.11) becomes negligible, $a_{\text{appr}}(t)$ thus has the right slope and $p_{\text{appr}}(T)$ equals $p(T)$ up to a factor (Fig. 2.3). We conclude that the duration distribution has a power-law tail with critical exponent 2. Expanding the exponential to third order yields a closed-form distribution that is a good approximation also for small T.

2.5 Simulations

We complement our analytics with simulations to (i) compare the avalanche distributions, (ii) exemplify the irrelevance of connectivity fluctuations, (iii) investigate the spatial spread of avalanches, (iv) address the robustness of the results, and (v) consider a typical experimental manipulation. If not stated otherwise, $N = 100$, $\tau = 10\,\text{ms}$ [57], $g = 500\,\text{Hz}$, $f_0 = 0.01\,\text{Hz}$, $f_{\text{sat}} = 2\,\text{Hz}$ [15, 40], somas are placed on unit square, $K^{-1} = 10^6\,\text{s}$ (quick growth, accelerating simulations) [15, 21, 22, 29]. The simulations use an event-based algorithm. Next spike times are determined using inverse transform sampling of the interspike-interval distribution; we avoid nonelementary functions by splitting each neuron's Poisson process into a homogeneous (rate f_0) and an inhomogeneous one.

An avalanche should be a sequence of offspring spikes of one spontaneous progenitor. To keep contact with the experimental literature, we analyze numerical data by binning time and considering spike sequences that are not separated by an empty bin as one avalanche [8, 58–60]. Our model yields analytical estimates for the probabilities that binning splits the first avalanche spikes or merges them with the next avalanche, as well as for splitting or merging an average avalanche. Keeping them moderate provides our bin sizes t_{bin} in terms of experimentally accessible quantities ($f_0, \tau, N, f_{\text{sat}}$), Appendix A.5. Results are robust against changing t_{bin}.

(i) In all simulations the model reaches a stationary state. The avalanche distributions agree well with the analytically derived ones, Fig. 2.3, the effects of binning and avalanche overlaps are small. We quantitatively test this agreement using the methods described in Refs. [61, 62]. For both size and duration distributions a pure power law is ruled out, as expected. For the size distribution, a power law with exponential cutoff, cf. Eq. (2.5), yields a good fit. The analytical values of the power-law

exponent, the cutoff $s_c(\sigma)$, and the resulting σ are closely matched.

(ii) The fluctuations of $\sum_j A_{ij}(t)$ and the deviations of $\bar{f}_i$ from f_{sat} are small (< 1‰, Fig. 2.2). Freezing the network ($K = 0$) in the stationary state has no effect on the avalanche statistics: Neuronal growth carries the system close to a critical point, but is not required later on. This is in agreement with self-organized criticality and excludes other mechanisms [63–65].

(iii) To investigate spatial spread near criticality, we compute covariances $C_{ij} = \langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle$ between numbers n_i, n_j of spikes contributed to single avalanches by neurons i, j with various distances. Covariances decay comparably slowly. Covariances and thus avalanches spread beyond direct connections, Fig. 2.4a.

(iv) To test robustness, we first freeze the growth in the stationary state and shuffle the output vectors (columns of the coupling matrix) between neurons. While this alters the network topology and breaks coupling symmetry, it leaves the essential total coupling strengths unchanged. Indeed, we observe little effect on avalanche sizes and durations. Second, we consider moderate nonadditivity of spike impacts. We introduce an absolute refractory period τ_{ref} after a sent spike, during which the neuron cannot spike again. We observe that although the refractory period limits the firing rate, the network reaches a stationary state with the same average individual rate f_{sat} as before: larger overlaps compensate refractoriness. For a refractory period about τ, which is often biologically plausible [39, 57], the statistics resemble the original one for small and medium size avalanches [Fig. 2.4b, red vs. gray]. Larger couplings and stop of avalanches lacking available neurons cause an excess of larger avalanches, followed by a strong reduction. Neurons still frequently contribute several spikes to avalanches. With long refractoriness, little similarity remains [Fig. 2.4b, green and blue; blue: our model with parameters adapted from Ref. [27], calcium variable present in Ref. [27] does not affect distribution shape].

(v) Manipulation of neural excitability or coupling strength via g causes subcriticality (decreased g) or excess of large avalanches (increased g), as in experiments [8, 66]. Our model predicts that the latter is balanced by network plasticity faster than the former, due to strongly increased activity, Appendix A.1.

2.6 Discussion and conclusion

We suggest an analytically tractable model for neural network growth, which may explain the emergence of subcritical and critical avalanche dynamics. It covers essential features of biological neurons such as operation in continuous time, spiking, leak currents, and network growth. Still, it allows the analytical computation of the avalanche size and duration distributions for subcritical and critical stationary states. Our numerical analysis confirms their validity and robustness and yields additional insight.

Two features are responsible for the emergence of the (near-)critical state (Fig. 2.3): (i) homeostatic growth to attain a saturation rate that is high compared to the spontaneous one (precise values of f_{sat} and f_0 are irrelevant), and (ii) linear summation of spike impacts. (i) implies that in the stationary state on average each spike generates nearly one successor. This holds for all networks with largely self-sustained activity. Usually, however, branching parameters vary, for example at high network activity spikes generate less offspring. This drives activity excursions back and generates non-power-law distributions [67, 68]. In our networks, (ii) implies that the branching parameter is the same for each spike. Small saturation rates yield subcritical dynamics (Fig. 2.3), strong nonlinearities other deviations [Fig. 2.4b]. Our model thus predicts that neural networks may develop criticality

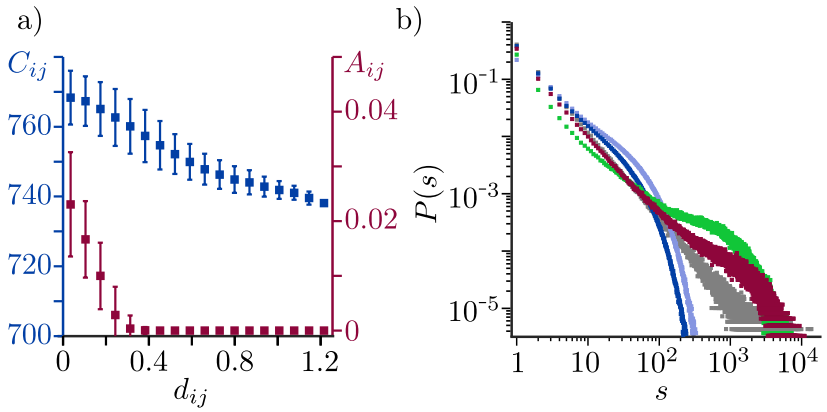

Figure 2.4: (a) Pairwise spike number covariances and overlaps as functions of the intersomatic distances d_{ij} (averages around a particular intersomatic distance, bars: standard deviations), $\sigma = 0.995$, $t_{\text{bin}} = 45$ ms. (b) Avalanche size distributions for the same model as in Fig. 3, but with absolute refractory period $\tau_{\text{ref}} = \tau$ (red), $\tau_{\text{ref}} = 4\tau$ (green), $\tau_{\text{ref}} = 0$ ms (gray) for reference, $t_{\text{bin}} = 45$ ms; $\tau = 5$ ms, $\tau_{\text{ref}} = 4\tau$, $f_0 = 0.1$ Hz, $f_{\text{sat}} = 0.8$ Hz, $t_{\text{bin}} = 10$ ms (blue, $t_{\text{bin}} = 45$ ms: light blue).

already due to their growth, that spontaneous spiking in such networks is low compared to saturation, and that spike effects add rather linearly and are independent of activity level. For example, starburst amacrine cells have radial dendritic trees, interact during development via dendro-dendritic excitatory connections and are reported to generate critical avalanches [14]. Our model predicts that higher precision measurements will reveal deviations as in Fig. 2.4b, due to the cells' long refractory periods.

Our network model is based on the neurobiologically more detailed ones [27, 29, 34]. Motivated by experiments, Ref. [34] proposes radial activity dependent neurite outgrowth steered by calcium dynamics and finds convergence to a stationary state for certain parameter ranges. To study avalanches, Ref. [27] adds stochastically spiking neurons, albeit with long refractoriness and larger f_0/f_{sat}, impeding analytical treatment and causing large deviations from criticality [Fig. 2.4b]; Refs. [29] assumes antagonistic growth of axons and dendrites and finds criticality, if a certain fraction of neurons becomes inhibitory; Refs. [45, 69, 70] consider more abstract homeostasis and neuron models.

Usually, models for neuronal avalanches only allow to estimate size and duration distributions numerically [14, 25, 27, 29, 30]. Reference [24] obtains an analytical expression of the size distribution for a discrete-time network. Our article derives size and duration distributions for a continuous-time spiking network model after self-organization. These distributions depend only on the experimentally accessible parameters f_0/f_{sat} and τ (duration scaling). The power-law exponents agree with experimentally found ones and those of simple branching processes [4, 6, 8, 14]. The duration distribution has power-law scaling at the tail [4, 6, 14], a fit to short avalanches [8] would yield different results. Our analytical expressions allow fast parameter scans, delineations of the (near-)critical regime and parameter estimations.

From a general perspective, avalanches in our model are clusters of a Hawkes process. While their size distribution can be straightforwardly computed [Eq. (2.4)], their duration distribution generally requires solving a nonlinear integral equation [55, 56]. Here we show that for Markovian Hawkes processes it follows from the solution of an ordinary differential equation [Eqs. (2.9),(2.10)] and we give closed-form approximations. This may find straightforward application in further fields of science where these processes are employed, for example, to characterize durations of financial market

fluctuations [54], earthquakes [71], violence [72], and epidemics [73, 74].

2.7 Acknowledgments

We thank Matthias Hennig, Anna Levina, Viola Priesemann, and Johannes Zierenberg for helpful discussions and the German Federal Ministry of Education and Research (BMBF) for support via the Bernstein Network (Bernstein Award 2014, 01GQ1501 and 01GQ1710).

References

[1] D. Marković and C. Gros, *Power laws and self-organized criticality in theory and nature*, Physics Reports **536** (2014) 41.

[2] P. Bak, C. Tang and K. Wiesenfeld, *Self-organized criticality - an explanation of 1/f noise*, Physical Review Letters **59** (1987) 381.

[3] P. Bak, C. Tang and K. Wiesenfeld, *Self-organized criticality*, Physical Review A **38** (1988) 364.

[4] S. Zapperi, K. Lauritsen and H. Stanley, *Self-organized branching processes: Mean field theory for avalanches*, Physical Review Letters **75** (1995) 4071.

[5] H. Jensen, *Self-organized criticality*, Cambridge: Cambridge Univ. Press, 1998.

[6] T. E. Harris, *The theory of branching processes*, Courier Corp., 2002.

[7] S. di Santo, P. Villegas, R. Burioni and M. A. Muñoz, *Simple unified view of branching process statistics: Random walks in balanced logarithmic potentials*, Physical Review E **95** (2017) 032115.

[8] J. Beggs and D. Plenz, *Neuronal avalanches in neocortical circuits*, Journal of Neuroscience **23** (2003) 11167.

[9] J. Beggs and D. Plenz, *Neuronal avalanches are diverse and precise activity patterns that are stable for many hours in cortical slice cultures*, Journal of Neuroscience **24** (2004) 5216.

[10] C. Haldeman and J. M. Beggs, *Critical Branching Captures Activity in Living Neural Networks and Maximizes the Number of Metastable States*, Physical Review Letters **94** (5 2005) 058101.

[11] W. L. Shew and D. Plenz, *The Functional Benefits of Criticality in the Cortex*, The Neuroscientist **19** (2013) 88.

[12] A. Mazzoni et al., *On the Dynamics of the Spontaneous Activity in Neuronal Networks*, PLoS One **2** (2007) e439.

[13] E. Gireesh and D. Plenz, *Neuronal avalanches organize as nested theta- and beta/gamma-oscillations during developent of cortical layer 2/3*, Proceedings of the National Academy of Sciences **105** (2008) 7576.

[14] M. H. Hennig, C. Adams, D. Willshaw and E. Sernagor, *Early-Stage Waves in the Retinal Network Emerge Close to a Critical State Transition between Local and Global Functional Connectivity*, Journal of Neuroscience **29** (2009) 1077.

[15] Y. Yada et al., *Development of neural population activity toward self-organized criticality.*, Neuroscience **343** (2017) 55.

[16] S. Kater and L. Mills, *Regulation of growth cone behavior via calcium*, Journal of Neuroscience **11** (1991) 891.

[17] S. Kater, M. Mattson, C. Cohan and J. Connor, *Calcium regulation of the neural growth cone*, Trends in Neurosciences **11** (1988) 315.

[18] C. Cohan and S. Kater, *Supression of neurite elongation and growth cone motility by electrical activity*, Science **232** (1986) 1638.

[19] F. van Huizen and H. Romijn, *Tetrodoxin enhances initial neurite outgrowth from fetal rate cerebral cortex cells in vitro*, Brain Research **408** (1987) 271.

[20] R. Fields, E. Neale and P. Nelson, *Effects of electrical activity on neurite outgrowth from mouse neurons*, Journal of Neuroscience **10** (1990) 2950.

[21] A. van Ooyen, *Using theoretical models to analyse neural development*, Nature Reviews Neuroscience **12** (2011) 311.

[22] G. Turrigiano, *Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function.*, Cold Spring Harbor Perspectives in Biology **4** (1 2012) a005736.

[23] A. van Ooyen and M. Butz-Ostendorf, *The rewiring brain*, London: Acad. Press, 2017.

[24] C. W. Eurich, J. M. Herrmann and U. A. Ernst, *Finite-size effects of avalanche dynamics.*, eng, Physical Review E **66** (2002) 066137.

[25] A. Levina, J. Herrmann and T. Geisel, *Dynamical synapses causing self-organized criticality in neural networks*, Nature Physics **3** (2007) 857.

[26] A. Levina, J. Herrmann and T. Geisel, *Phase transitions towards criticality in a neural system with adaptive interactions*, Physical Review Letters **102** (2009) 118110.

[27] L. Abbott and R. Rohrkemper, *A simple growth model constructs critical avalanche networks*, Progress in Brain Research **165** (2007) 13.

[28] V. Gómez, A. Kaltenbrunner, V. López and H. Kappen, "Self-organization using synaptic plasticity", *Advances in Neural Information Processing Systems 21*, Curran Associates, Inc., 2009 513.

[29] C. Tetzlaff, S. Okujeni, U. Egert, F. Wörgötter and M. Butz, *Self-organized criticality in developing neural networks*, PLoS Computational Biology **6** (2010) e100103.

[30] B. Del Papa, V. Priesemann and J. Triesch, *Criticality meets learning: Criticality signatures in a self-organizing recurrent neural network*, PLoS One **12** (2017) 1.

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

[32] S. Bornholdt and T. Röhl, *Self-organized critical neural networks*,
Physical Review E **67** (2003) 066118.

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

[34] A. van Ooyen and J. van Pelt, *Activity dependent outgrowth of neurons and overshoot phenomena in developing neural networks*, Journal of Theoretical Biology **167** (1994) 27.

[35] W. Gerstner and W. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*,
Cambridge: Cambridge Univ. Press, 2002.

[36] R. Kempter, W. Gerstner and J. L. Van Hemmen, *Hebbian learning and spiking neurons*,
Physical Review E **59** (1999) 4498.

[37] A. Burkitt, M. Gilson and J. van Hemmen,
Spike-timing-dependent plasticity for neurons with recurrent connection,
Biological Cybernetics **96** (2007) 533.

[38] V. Pernice, B. Staude, S. Cardanobile and S. Rotter,
How structure determines correlations in neuronal networks,
PLoS Computational Biology **7** (2011) e1002059.

[39] C. Koch, *Biophysics of computation*, Oxford: Oxford Univ. Press, 1999.

[40] K. J. Ford, A. L. Félix and M. B. Feller,
Cellular Mechanisms Underlying Spatiotemporal Features of Cholinergic Retinal Waves,
Journal of Neuroscience **32** (2012) 850.

[41] A. G. Hawkes, *Spectra of Some Self-Exciting and Mutually Exciting Point Processes*,
Biometrika **58** (1971) 83.

[42] J. Barral and A. D Reyes, *Synaptic scaling rule preserves excitatory–inhibitory balance and salient neuronal network dynamics*, Nature Neuroscience **19** (2016) 1690.

[43] M. Abeles, *Corticonics: Neural circuits of the cerebral cortex*, Cambridge Univ. Press, 1991.

[44] A. van Ooyen et al., *Independently outgrowing neurons and geometry-based synapse formation produce networks with realistic synaptic connectivity.*, PloS One **9** (1 2014) e85858.

[45] J. Zierenberg, J. Wilting and V. Priesemann,
Homeostatic Plasticity and External Input Shape Neural Network Dynamics,
Physical Review X **8** (3 2018) 031018.

[46] K. S. Crump and C. J. Mode, *A general age-dependent branching process. I*,
Journal of Mathematical Analysis and Applications **24** (1968) 494.

[47] J. C. Tanner, *A Derivation of the Borel Distribution*, Biometrika **48** (1961) 222.

[48] V. Priesemann et al., *Spike avalanches in vivo suggest a driven, slightly subcritical brain state.*,
Frontiers in Systems Neuroscience **8** (2014) 108.

[49] M. Abramowitz, I. A. Stegun et al.,
Handbook of mathematical functions with formulas, graphs, and mathematical tables, vol. 9,
Dover, New York, 1972.

[50] D. Oakes, *The Markovian self-exciting process*, Journal of Applied Probability **12** (1975) 69.

[51] E. Bacry, I. Mastromatteo and J.-F. Muzy, *Hawkes processes in finance*,
Market Microstructure and Liquidity **1** (2015) 1550005.

[52] X. Gao and L. Zhu,
"Limit theorems for Markovian Hawkes processes with a large initial intensity",
Stoch. Process. Appl., in press, 2018.

[53] A. Dassios and H. Zhao, *A dynamic contagion process*,
Advances in Applied Probability **43** (2011) 814.

[54] E. Errais, K. Giesecke and L. R. Goldberg, *Affine point processes and portfolio credit risk*,
SIAM Journal on Financial Mathematics **1** (2010) 642.

[55] A. G. Hawkes and D. Oakes, *A cluster process representation of a self-exciting process*,
Journal of Applied Probability **11** (1974) 493.

[56] J. Møller and J. G. Rasmussen, *Perfect Simulation of Hawkes Processes*,
Advances in Applied Probability **37** (2005) 629.

[57] P. Dayan and L. Abbott,
Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems,
Cambridge: MIT Press, 2001.

[58] V. Priesemann, M. H. Munk and M. Wibral,
Subsampling effects in neuronal avalanche distributions recorded in vivo,
BMC Neuroscience **10** (2009) 40.

[59] G. Hahn et al., *Neuronal avalanches in spontaneous activity in vivo.*,
Journal of Neurophysiology **104** (6 2010) 3312.

[60] A. Levina and V. Priesemann, *Subsampling scaling.*, Nature Communications **8** (2017) 15140.

[61] A. Clauset, C. R. Shalizi and M. E. J. Newman, *Power-Law Distributions in Empirical Data*,
SIAM Review **51** (2009) 661.

[62] S. V. Aksenov et al., *Application of the combined nonlinear-condensation transformation to problems in statistical analysis and theoretical physics*,
Computer Physics Communications **150** (2003) 1.

[63] Didier Sornette, *Sweeping of an instability : an alternative to self-organized criticality to get powerlaws without parameter tuning*, Journal de Physique I **4** (1994) 209.

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

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

[66] W. L. Shew, H. Yang, S. Yu, R. Roy and D. Plenz, *Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches*, Journal of Neuroscience **31** (2011) 55.

[67] T. Vogels and L. Abbott, *Signal Propagation and Logic Gating in Networks of Integrate-and-Fire Neurons*, Journal of Neuroscience **25** (2005) 10786.

[68] A. Kumar, S. Schrader, A. Aertsen and S. Rotter, *The High-Conductance State of Cortical Networks*, Neural Computation **20** (2008) 1.

[69] A. Levina, U. Ernst and J. M. Herrmann, *Criticality of avalanche dynamics in adaptive recurrent networks*, Neurocomputing **70** (2007) 1877.

[70] F. Droste, A.-L. Do and T. Gross, *Analytical investigation of self-organized criticality in neural networks*, Journal of The Royal Society Interface **10** (2013) 20120558.

[71] T. Wang, M. Bebbington and D. Harte, *Markov-modulated Hawkes process with stepwise decay*, Annals of the Institute of Statistical Mathematics **64** (2012) 521.

[72] E. Lewis, G. Mohler, P. J. Brantingham and A. L. Bertozzi, *Self-exciting point process models of civilian deaths in Iraq*, Security Journal **25** (2012) 244.

[73] H. Kim, *Spatio-temporal point process models for the spread of avian influenza virus (H5N1)*, 2011, URL: `http://www.escholarship.org/uc/item/8nc0r19n`.

[74] F. Schoenberg, M. Hoffmann and R. Harrigan, *A recursive point process model for infectious diseases*, 2017, eprint: `arXiv:1703.08202`.

CHAPTER 3

Drifting Assemblies for Persistent Memory: Neuron Transitions and Unsupervised Compensation

Change is ubiquitous in living beings. In particular, the connectome and neural representations can change. Nevertheless behaviors and memories often persist over long times. In a standard model, associative memories are represented by assemblies of strongly interconnected neurons. For faithful storage these assemblies are assumed to consist of the same neurons over time. Here we propose a contrasting memory model with complete temporal remodeling of assemblies, based on experimentally observed changes of synapses and neural representations. The assemblies drift freely as noisy autonomous network activity or spontaneous synaptic turnover induce neuron exchange. The gradual exchange allows activity-dependent and homeostatic plasticity to conserve the representational structure and keep inputs, outputs and assemblies consistent. This leads to persistent memory. Our findings explain recent experimental results on temporal evolution of fear memory representations and suggest that memory systems need to be understood in their completeness as individual parts may constantly change.

This chapter is a reproduction with minor alterations of the article of the same title that was published in Proceedings of the National Academy of Sciences under the reference: Yaroslav Felipe Kalle Kossio, Sven Goedeke, Christian Klos, Raoul-Martin Memmesheimer, Proc. Natl. Acad. Sci. U.S.A., Nov 2021, 118 (46) e2023832118, https://doi.org/10.1073/pnas.2023832118. The supplementary material is reproduced in Appendix B.

3.1 Introduction

Organisms change over time, on many different levels. This holds in particular for the synapses in neural networks [1]: They change their impact and also appear and vanish. On the one hand, this weight and structural plasticity is activity dependent. Such forms have been argued and directly shown to be crucial for learning [2]. On the other hand, weight changes and turnover of connections with

similar magnitude occur spontaneously, in excitatory and inhibitory synapes, independent of previous spiking activity and in its absence [3–7]. A similar dichotomy exists for neural representations: They change due to adaptive learning in order to improve task performance, but also spontaneously, often without affecting behavior. The latter has been observed in areas storing long-term memories [8], in sensory areas, for place cells, location and goal-selective cells, and in motor areas [9, 10]. The changes over the durations of the experiments were mostly only partial.

Environments change as well. To flexibly adapt, higher animals acquire information and retain it by forming memories in the brain. In a widely used model, a memory is represented by one or several (depending on their definition) neuronal assemblies, ensembles of strongly interconnected neurons [11, 12]. If an assembly is partially excited, for example by an external input, the remainder of the neurons follow, leading to associative memory recall. For faithful memory storage the ensemble of neurons forming an assembly is assumed to remain the same [13]. Previous theoretical analysis has carefully studied the formation and maintenance of such static neuronal assemblies [14–22]. In particular, it has been suggested that in presence of noisy autonomous (without receiving external stimulation or feedback) network activity [15–18] and spontaneous (activity-independent) synaptic changes [21, 23], assemblies are preserved with the help of activity-dependent synaptic plasticity.

Based on the experimentally observed changes of synaptic weights and connections and neural representations we develop a contrasting associative memory model where assemblies are ever and completely changing; they drift or "swim". This happens gradually, by successive exchange of individual neurons. The neuron ensembles forming the same assembly at distant times are not directly related, but indirectly via the ensembles forming the assembly at the times in between. Using an analogy of ref. [24] (Appendix B.2), this is comparable to a thread, which consists of many rather short overlapping fibers; the ensembles of fibers in spatially distant parts are not directly related. In our model, the participation of single neurons in the memory representation overlaps, Fig. 3.1a, like the participation of fibers in the thread. As a consequence, viewed over time the representation looks like a continuous thread, Fig. 3.1b. While fibers adhere together due to the friction between them, neurons in the assembly adhere due to increased synaptic weights. The inputs and outputs track the course of the "assembly thread" to keep behavior and memory stable, Fig. 3.1c; they connect at each time to the correct ensemble of neurons that currently forms the required neural representation. Stable input neurons may be located in the sensory periphery, but also within the brain [9], for example in the primary visual cortex and the dentate gyrus, the input area of the hippocampus [25]; motor neurons are candidates for stable output neurons. We will refer to both input and output neurons (Fig. 3.1c) as periphery neurons and to the assembly forming neurons (Fig. 3.1a) as interior ones.

We demonstrate the feasibility of our memory model using neural networks at different levels of complexity and with different types of dynamics. Numerical simulations and theoretical analysis reveal that assembly drift can be driven by synaptic weight fluctuations due to noisy autonomous activity or by spontaneous synaptic turnover (activity-independent appearance and disappearance of synaptic connections). The overall representational structure and memory are maintained by activity-dependent and homeostatic synaptic plasticity. Furthermore, we find that assembly drift can be directly related to the evolution of fear memory representations uncovered in recent experiments [8] and that drifting assemblies are suitable for computation.

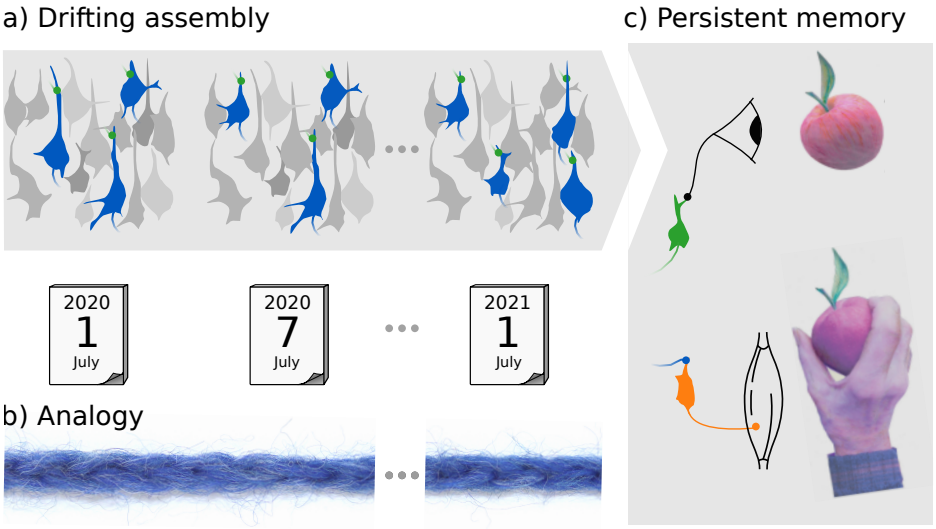

Figure 3.1: Assembly drift and persistent memory. (a) At two nearby times a similar ensemble of neurons forms the neural representation of, for example, "apple" (compare the blue colored assembly neurons at the first and the second time point). At distant times the representation consists of completely different ensembles (blue colored assembly neurons at the first and the third time point). Due to their gradual change, temporally distant representations are indirectly related via ensembles in the time period between them. (b) Parts of a thread possess the same form of indirect relation: Nearby parts are composed of similar ensembles of fibers, while distant ones consist of different ensembles, which are connected by those in between. (c) The complete change of memory representations still allows for stable behavior. In the figure, a tasty apple is perceived. At different times, this triggers different ensembles that presently form the representation of "apple", see (a). Assembly activation initiates a reaching movement towards the apple, despite the dissimilarity of the activated neuron ensembles. Memory and behavior are conserved because the gradual change of assembly neurons enables the inputs (green) and outputs (orange) to track the neural representation.

3.2 Results

A spiking neural network model for drifting memory representations

We show the feasibility of our concept (Figs. 3.1, 3.2a) using networks of leaky integrate-and-fire (LIF) model neurons. Figure 3.2b displays the matrix of synaptic weights between the excitatory neurons (henceforth simply: weight matrix) of such a network with 90 interior and 12 periphery neurons. (Synapses from, to and between inhibitory neurons are modeled as uniform and static throughout the article.) The network is initialized with three assemblies; each has two input and two output neurons. The strength of a synapse from neuron j to neuron i is given by the entry w_{ij} of the weight matrix; we measure it in terms of the peak excitatory postsynaptic potential (EPSP, units: mV). The synaptic weights change due to spike timing dependent plasticity (STDP) with symmetric window and due to divisive homeostatic normalization, which ensures that both input and output weights sum to w_{sum} for each neuron (Methods, Appendix Fig. B.1). In addition, they are restricted by a maximal possible weight w_{max}. Periphery neurons differ from interior neurons on the "physiological level" only by a larger w_{max}, smaller w_{sum} and weaker STDP (smaller amplitude). We further assume that periphery neurons do not connect to each other, because they might lie in distinct brain areas with little interconnectivity.

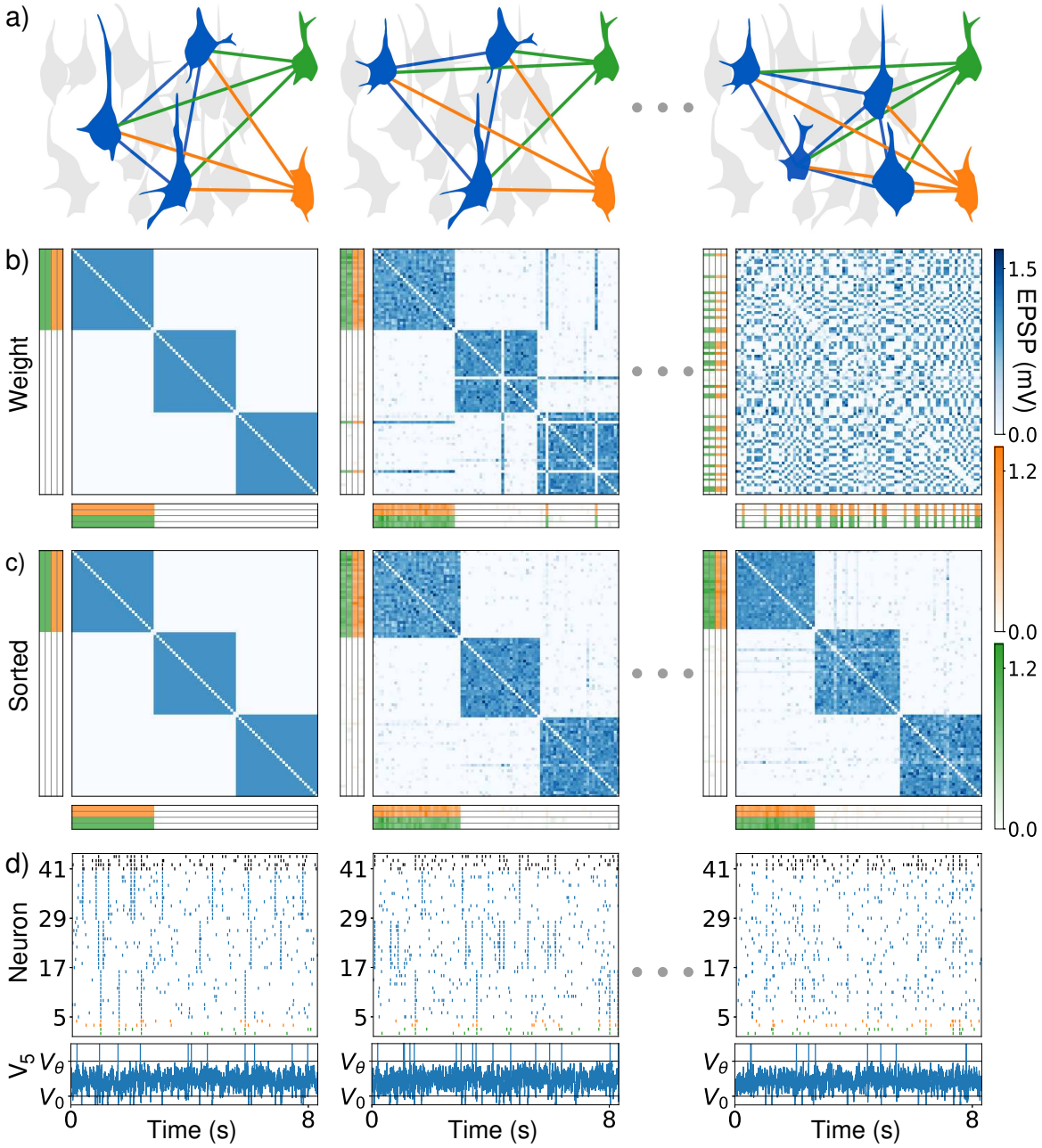

Figure 3.2: Drifting assemblies in spiking neural networks. (a) Schematics emphasizing strong synaptic coupling. While an assembly drifts freely (blue colored assembly neurons) within the interior neurons, its input and output neurons (green and orange) follow it by adapting their synaptic weights. (b) Weights between interior neurons (blue weight matrix), from input and output neurons to the interior neurons (green and orange vertical weight matrices) and from the interior to the input and output neurons (green and orange horizontal weight matrices). Input (output) weights of neuron i are displayed as ith row (column). Only weights of the four periphery neurons initially (and thus for all times) attached to assembly 1 are shown for clarity. Caption continues on next page.

Figure 3.2: First column: Network initialization with three assemblies. Second column, after 27 minutes: Noisy autonomous spiking activity has already driven several interior neurons to attach to a new assembly (blue weight matrix, horizontal and vertical "lines" indicating the changed input and output preference). Third column, after 30 hours (h): The assemblies have drifted away, the weight matrix is completely remodeled. (c) Like (b) but with neurons reordered according to assemblies that they belong to, using a clustering algorithm. The assemblies remain intact and the periphery neurons stay strongly coupled to assembly 1. (d, upper) Spike trains of the input (green) and output (orange) neurons of assembly 1, of 12 neurons from each of the ensembles that initially form assembly 1 (5-16), 2 (17-28) and 3 (29-40) and of four inhibitory neurons (black). (d, lower) Membrane potential of the first interior neuron fluctuates irregularly. Spikes are marked by vertical lines above threshold V_θ, reset is to V_0.

In Fig. 3.2b the neurons initially forming assembly 1 are displayed with lowest indices, in the upper left corner of the weight matrix. Periphery neurons with strong input from and output to assembly 1 therefore have strong weights at the left part of the horizontal and at the upper part of the vertical weight matrices. The interior neurons forming the assembly then gradually change, but the assembly is preserved: it drifts freely in the network. Furthermore its input and output neurons stay the same, Fig. 3.2c. This holds for all assemblies.

As part of autonomous network activity, neurons forming an assembly occasionally spike synchronously, Fig. 3.2d. Such reactivations appear at later times dispersed over the neuron indices, since the indexing does not fit the assembly structure anymore. Background spiking is irregular and asynchronous; the membrane potentials of neurons fluctuate irregularly.

We verified that the drifting assemblies have the associative memory property of activating after being partially excited and that they are functional in the sense that they mediate an input-output association: after a sufficiently strong stimulation of its input neurons, an assembly activates and stimulates its output neurons to generate increased spiking activity (Appendix Fig. B.2). We note that drifting assemblies can also occur in network models without periphery neurons (Appendix Fig. B.3).

Noisy autonomous activity gives rise to drifting assemblies

The assembly drift is on the level of single interior neurons reflected by characteristic dynamics: comparably long times of stable assembly membership, which are interspersed with fast switches between them, Fig. 3.3a. Periphery neurons do not switch assemblies, Fig. 3.3b.

What is the mechanism underlying switching? A neuron generally spikes when its assembly reactivates, resulting in strengthened synaptic coupling between them due to STDP. Coincident spiking of an interior neuron together with reactivations of another assembly can increase the synaptic weights between that assembly and the neuron via STDP, Fig. 3.3c. The homeostatic competition between synapses then leads to weakening of the synapses between the neuron and its current assembly. If the weight perturbation is not eliminated (for example by spiking of the neuron together with its current assembly), the neuron spikes with higher probability together with the other assembly, which results in even stronger binding. Therefore, a neuron sometimes tends to leave its current and switch to another assembly. There are further mechanisms that contribute to switching in our networks: If an assembly reactivates without near simultaneous spiking of the neuron, the homeostatic normalization together with the strengthening of the synapses between the reactivated neurons leads to weakening of the synapses between the neuron and the assembly. In addition, sometimes moderately asynchronous spiking of the neuron with respect to an assembly reactivation results in weakening

of their interconnecting synapses due to STDP. Finally, smaller weight changes are induced by the background activity. In the transition phase, when input synapses from both assemblies are strong, the weights are most volatile. For one of its neurons, an assembly can thus be thought of acting like a potential and noise well, Fig. 3.3d. A neuron jumps back towards the bottom after a perturbation of its weights, but a number of strong perturbations in a short period of time can induce a transition from one well to another. Due to weaker STDP periphery neurons experience smaller perturbations, which are insufficient to escape the well.

Neuron switching and thus assembly drift continues over time. Weights and assemblies completely and continuously remodel, Fig. 3.3e-g. We note that neurons often generate transiently strong connections to another than their current assembly (down- and upstrokes in Fig. 3.3a). At these times and during transitions they are to some degree shared between two assemblies and generate an overlap between them.

The mechanisms of assembly stabilization that ensure in our networks long times of assembly membership have been previously described for static assemblies: Correlated activity, e.g. due to reactivations, leads to strengthening of internal and weakening of inter-assembly synaptic weights via long term potentiation (LTP) and depression (LTD) [16–19, 23, 26]. Additionally, homeostatic plasticity introduces competition, which weakens inter-assembly synapses, as they usually get potentiated less by activity-dependent plasticity [27–29].

While in our networks, the interior neurons switch between assemblies, Fig. 3.3a, no assembly vanishes. Responsible for this is likely a higher reactivation rate of smaller assemblies due to stronger internal synapses (Appendix Fig. B.4): More frequent reactivation leads to stronger binding and recruitment of neurons. Since the assemblies compete for neurons (cf. also [30]), smaller than average assemblies have the tendency to grow and larger ones to shrink. We note that, depending on the network parameters, assemblies can also emerge spontaneously for random initial weights (Appendix Fig. B.5), as observed for static assemblies [18, 22, 28].

Simplified models of neuron switching and assembly drift

As a prerequisite for our first effective model of neuron switching and to further elucidate the mechanism of the neuron transitions, we examine the change Δw_1 of the summed input weight w_1 from assembly 1 to a neuron, Fig. 3.4a. We focus on a network with two assemblies and without periphery neurons for simplicity (see Appendix Fig. B.8 for the same analysis in a network with three assemblies). We consider only the inputs, since the influence of a single neuron's output on an assembly is small. Weights are normalized, such that the total summed input equals 1 and thus w_1 is between 0 and 1. We consider all instances where the weight w_1 has a certain value, measure the ensuing changes Δw_1 during the typical inter-spike-interval and compute their average $\overline{\Delta w_1}(w_1)$ and standard deviation $\mathrm{Std}(\Delta w_1)(w_1)$ (Methods). If $\overline{\Delta w_1}(w_1)$ is greater than zero, the input strength from assembly 1 on average increases: The neuron is "drawn towards" this assembly. This happens for w_1 close to 1, Fig. 3.4b left, i.e. when the neuron is part of assembly 1. We note that the actual changes are finite size jumps. Since we consider two assemblies, the cases $w_1 \approx 0$ and $w_1 \approx 1$ are symmetric. $w_1 \approx 0$ means that the neuron belongs to assembly 2 and tends to stay away from assembly 1.

The analysis shows that we can understand the switching dynamics in the networks of LIF neurons, Fig. 3.4a left column, as a random walk between the wells: in each step there is a deterministic update $\overline{\Delta w_1}(w_1)$ and a fluctuation of size $\mathrm{Std}(\Delta w_1)(w_1)$ around it (Appendix B.5). The dynamics may be visualized by a potential $U(w_1)$ (Methods), where the neuron behaves roughly like an object

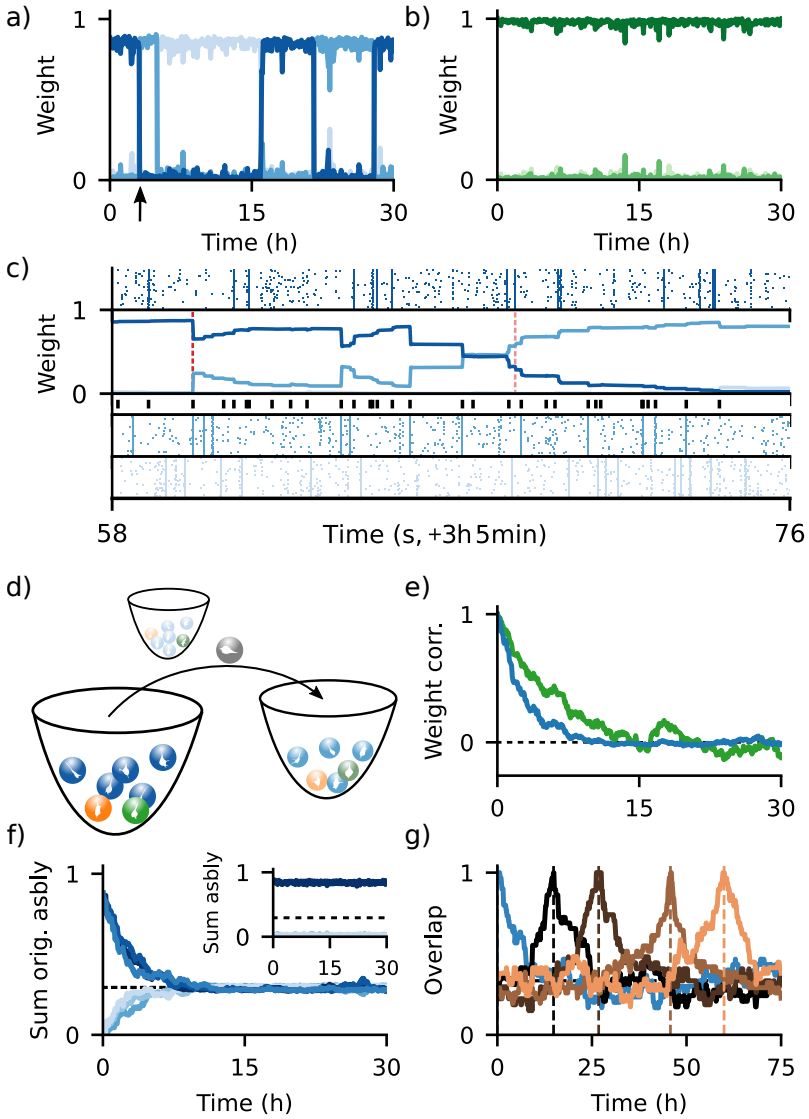

Figure 3.3: Analysis of drifting assemblies and their periphery neurons. (a) Switching of interior neurons in a simulation. Normalized summed weights between a neuron and current assemblies 1, 2 and 3 (dark to light blue) show temporary membership and fast transitions. (b) Periphery neurons stay attached to their assembly. Display like (a), with greens indicating periphery neuron-assembly weights. (c) Close up of the switching event indicated in (a) by an arrow. Raster plots: spikes of neurons in the three assemblies. The switching neuron is in assembly 1 (first subpanel) before and in assembly 2 after its transition. Second subpanel: total weight between each assembly and the neuron. Red dashed: switching neuron spikes together with reactivation of assembly 2; light red: failure to spike with assembly 1.Third subpanel: Spikes of the switching neuron. (d) Schematic illustration of the mechanism underlying assembly drift. Noise drives balls (neurons) out of wells, which are generated by the different assemblies. They move to other wells (neurons switch assemblies). Periphery neurons (green, orange) experience too little noise to be pushed out of the wells; they stick with their assemblies. (e) Complete network weight remodeling. Pearson correlation between initial and later weights of interior (blue) and periphery-interior synapses (green) converge to chance level (black, 0). Caption continues on next page.

Figure 3.3: (f) Complete assembly remodeling. Summed weights within and between (darker and lighter blues) the three initial assemblies converge to chance level (1/3 of recurrent interior coupling, black). Inset: Maintenance of representational structure. Sum of weights within current assembly 1 (dark blue) and between it and the current other two assemblies (light blues). (g) Assembly drift continues over time. Overlap between the neuron ensemble forming assembly 1 at a reference time, with the neuron ensembles forming assembly 1 at other times. Reference times (dashed verticals) are: initialization (blue curve) and first to fourth complete remodeling time (dark to light brown curves). Chance level is 1/3 (black, mostly covered).

jumping down the potential's slope. We see that the average weight change alone lets the neuron stay in the wells near $w_1 \approx 1$ or $w_1 \approx 0$, Fig. 3.4c left. The noise in the weight changes, quantified by $\mathrm{Std}(\Delta w_1)(w_1)$ in Fig. 3.4d left, is thus crucial for transitions, like in noise-activated transitions between meta-stable states [31]. A modified model without the deterministic update shows that already the presence of stronger noise in the transition zone prevents the neuron from lingering there and lets it choose one of the assemblies (Appendix B.5). In other words: already the inhomogeneous noise alone would lead to the observed meta-stable states (noise-induced multistability) [32].

We also obtain an effective random walk model for switching from first principles, Fig. 3.4 middle column. It assumes that neurons spike and assemblies reactivate according to Poisson processes with fixed rates. A "test neuron" further spikes with input weight dependent probability during assembly reactivations. We can then compute the input weight changes of the test neuron resulting from co-spiking with its own assembly, coincident spiking with the other assembly and background spiking (Methods, Appendix B.6). The model generates switching behavior, potentials and noise similar to LIF networks, confirming our previous explanations for them. In particular, the model confirms that co-spiking with the neuron's own assembly generates potential wells and shows that stronger weight fluctuations in the transition zone result from opposing weight changes evoked by similarly frequent co-spiking with both assemblies.

To enable long-term simulations for a comparison with experiments, we consider a binary model with weight-dependent covariance plasticity rule (Methods), which also shows assembly drift due to noisy autonomous activity (Appendix Fig. B.6). Its switching dynamics differs slightly from those of the LIF model, as there is an additional potential well in the middle, Fig. 3.4 right. The strong noise in this region, however, prevents the neuron from getting trapped there. Also for the binary network, the switching dynamics can be reduced to an effective random walk (Methods, Appendix Fig. B.9). This reduced model shows that the additional potential well occurs because the neuron co-spikes in the transition zone with both assemblies with equal probability and homeostatic normalization favors the assembly with smaller weight, preventing the emergence of selectivity.

Spontaneous synaptic turnover also gives rise to drifting assemblies

Assembly drift can also be driven by activity-independent synaptic changes. To show this, we introduce spontaneous turnover of connections, i.e. synapses appear and disappear [3]. For simplicity, we assume a uniform turnover rate that does not depend on the current synaptic weight. A synaptic connection between two excitatory neurons in our networks has a finite expected lifetime (Methods). Similarly, if the synapse is absent, it has a finite average absence time until it reappears, with initial weight zero. The excitatory connectivity thus completely remodels. The average lifetime of a synapse is about half an hour. Experiments indicate average lifetimes in absence of activity that range from minutes

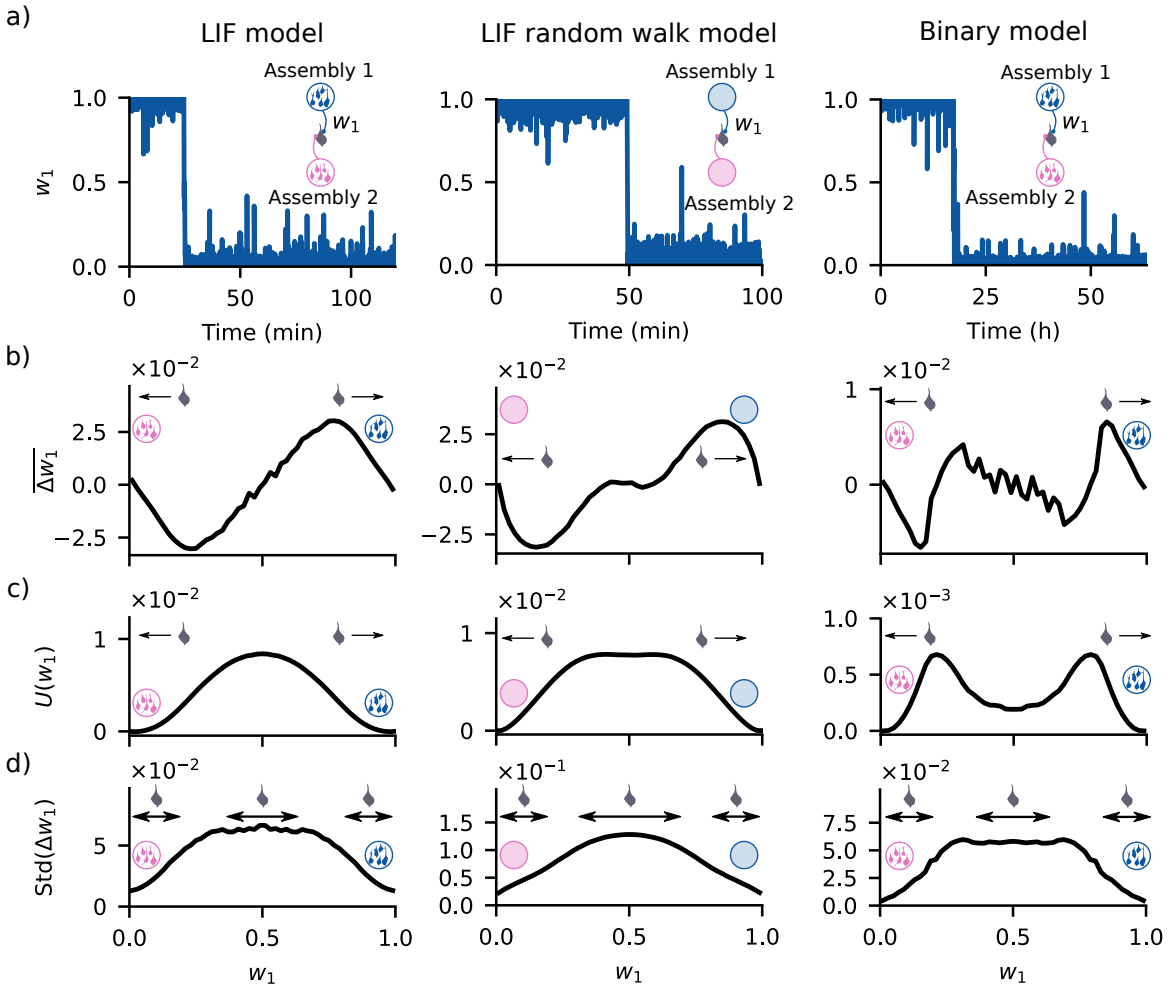

Figure 3.4: Mechanisms of neuron transitions between assemblies driven by weight changes due to noisy autonomous activity. The mechanisms for the LIF model (left column) are well captured by a random walk model (middle); the mechanisms of the binary model (right) are similar. (a) Long-term membership of a neuron in assemblies and fast transition between them. Large summed input weight w_1 from assembly 1 to the neuron reflects membership in this assembly. (b) Average change $\overline{\Delta w_1}$ of the summed input weight w_1 from assembly 1, as a function of the current weight value. (c) Corresponding potential $U(w_1)$ for the average weight changes. The average weight changes are roughly similar to the displacement of an object jumping down the landscape given by $U(w_1)$. The two wells near 0 and 1 induce meta-stable states corresponding to the different assembly memberships. (d) Standard deviation Std(Δw_1) of the change of the summed input weight w_1 from assembly 1, as a function of the current weight value.

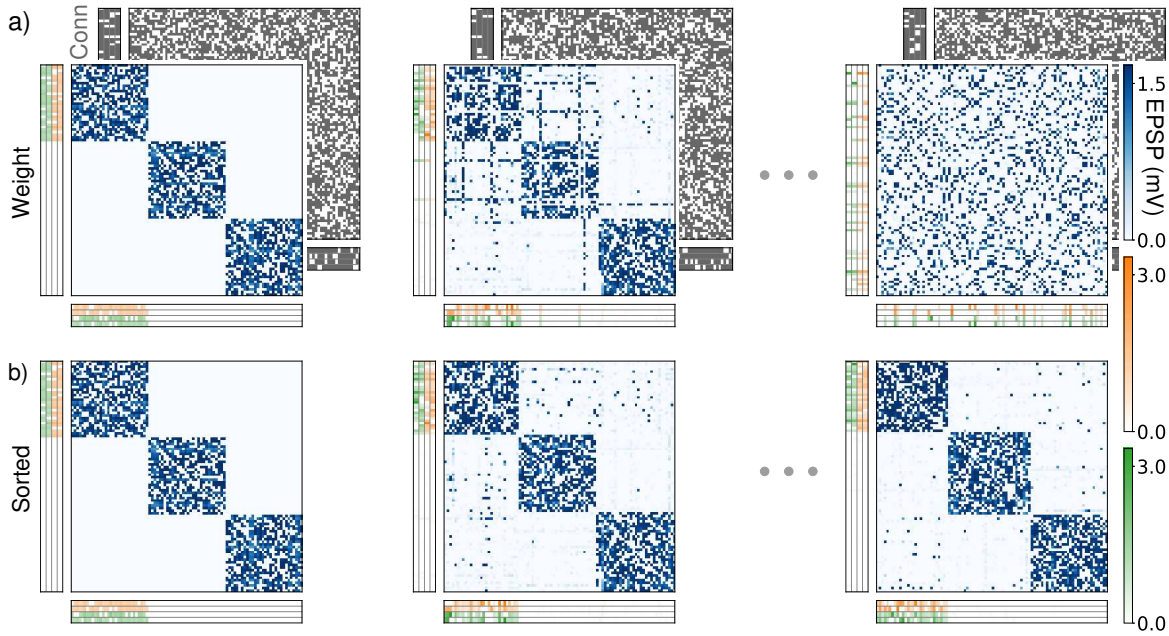

Figure 3.5: Spontaneous synaptic turnover gives rise to assembly drift. (a) The matrix in the background shows the incomplete and spontaneously changing connectivity that drives the drift (gray: present synapses). First column shows the weight and connectivity matrices after initialization, second column after an hour and third column after 37 h. Otherwise the depiction is like in Fig. 3.2b. (b) like Fig. 3.2c.

for immature synapses to two months for mature ones with large weights [3, 6]. We choose shorter lifetimes to reduce the simulation time to a feasible amount and because in biological networks there are also activity-independent weight fluctuations [5]. The presence or absence of the synapse from neuron j to i is indicated by a one or zero in the entry c_{ij} of the connectivity matrix of the network.

Fig. 3.5 displays assembly drift in such a network of LIF neurons (Methods). For the employed parameters, the spontaneous synaptic turnover drives the drift: Without it, assemblies stay invariant after a short time of adaptation to the fixed connectivity matrix (Appendix Fig. B.10), because the weight changes due to STDP are weaker than in the network of Fig. 3.2 and noisy activity alone is insufficient to drive switching of neurons. Analysis of switching events (Appendix Fig. B.11) reveals that strong decreases of the total number of available input synapses from its current assembly precede a neuron's switch. This indicates that in networks with spontaneous synaptic turnover switching neurons are "pushed out" of their assembly by occasional spontaneous downward fluctuations of the number of available input connections from their assembly. If the number of input connections from its assembly becomes too small, homeostasis lets a neuron strengthen its synapses from other assemblies to maintain the desired input level. Supported by weight fluctuations due to noisy autonomous activity, the neuron will then start to spike together with other assemblies and finally perform a fast transition to one of them. This last part of the switching dynamics is thus similar to the one described above. In other words, downward fluctuations of the number of available input connections bring an interior neuron so close to the edge of its current potential and noise well (assembly) that the otherwise insufficient weight fluctuations are able to induce a transition to another one. Complete remodeling of the weights occurs on a longer timescale than spontaneous synaptic turnover (Appendix Fig. B.11),

since STDP and homeostatic competition compensate large parts of the turnover. A periphery neuron in the network of Fig. 3.5 differs from an interior neuron by higher average connectivity and, as for our networks where (only) noisy activity drives assembly drift, by smaller w_{sum} and larger w_{max} (Methods). It can therefore compensate a higher number of lost input connections from its assembly by increasing the weights of the remaining ones.

A model for experimentally found evolution of fear memory representations

Change of a long-term memory representation was recently experimentally observed in the prelimbic cortex [8], a structure known to be crucial for context dependent aversive memory [33]. In the immunostaining experiment different groups of mice underwent context dependent fear conditioning with an auditory stimulus. This was followed by two retrievals, the first after 1, 7 or 14 days, the second after 28 days (Fig. 3.6a). Neurons active during fear conditioning or neurons active during the first retrieval were detected by using a genetic marker, TRAP2. Additionally neurons active during the second retrieval were detected by Fos expression. The numbers of neurons forming the memory representation were approximately the same in all sessions, as well as the behavioral response (freezing) during retrievals. In contrast, the subset of neurons labeled by both markers and thus the overlap of the neuron ensembles forming the memory representation, increased with decreasing time between the two labeling sessions.

We conjectured that the experimental findings can be explained if we assume that a drifting assembly forms the memory representation. To model the system, we combine an assembly in prelimbic cortex with "context" input neurons in the hippocampus and auditory stimulus "tone" input neurons in the auditory cortex. These neurons signal that the animal is in the cage where fear conditioning had taken place and that the conditioned tone is played. The assembly's output neurons are in the amygdala [34], Fig. 3.6b (Methods). We take the probability of their activation as indicator of the animal's freezing time. In an animal, their activation is linked to the freezing response via internal amygdala processing and downstream areas, which we assume to limit the freezing response to the observed 60% of the time and to ensure that it occurs in continuous, longer stretches [33].

The model replicates the main experimental findings (Fig. 3.6c left): The overlap between the neural representation of memory on day 28 and an earlier session increases for shorter periods between sessions (Fig. 3.6c right), because during shorter periods less assembly neurons switch. The assembly sizes are on average unchanged, as the assemblies only drift. Finally, the behavioral response is conserved, because the input and output neurons faithfully follow the drift. Behavior was also conserved in experiments (Supplementary Fig. 4a in [8]).

The model also accounts for more details of the experimental results. The conditioned tone alone induced more freezing than the conditioned context alone. The model replicates this (Fig. 3.6d,e right, dashed lines), since we chose the tone in auditory cortex to be represented by more input neurons than the context in hippocampus. The experiments included photostimulation of neurons that were labeled during fear conditioning or the first of the two retrieval sessions. In absence of the conditioned tone, this affected behavior the stronger the smaller the period between retrieval and stimulation was, Fig. 3.6d,e left, solid lines. Our model captures this result, since the more time has passed since labeling, the more of the labeled neurons have transitioned away from the assembly. The laser thus stimulates a smaller part of the current assembly and activates it more rarely, which is reflected in the output, Fig. 3.6d,e right, solid lines. In absence of photostimulation, the level of representation remodeling has no impact on the behavioral output (Supplementary Fig. 4a in [8]). Our model explains

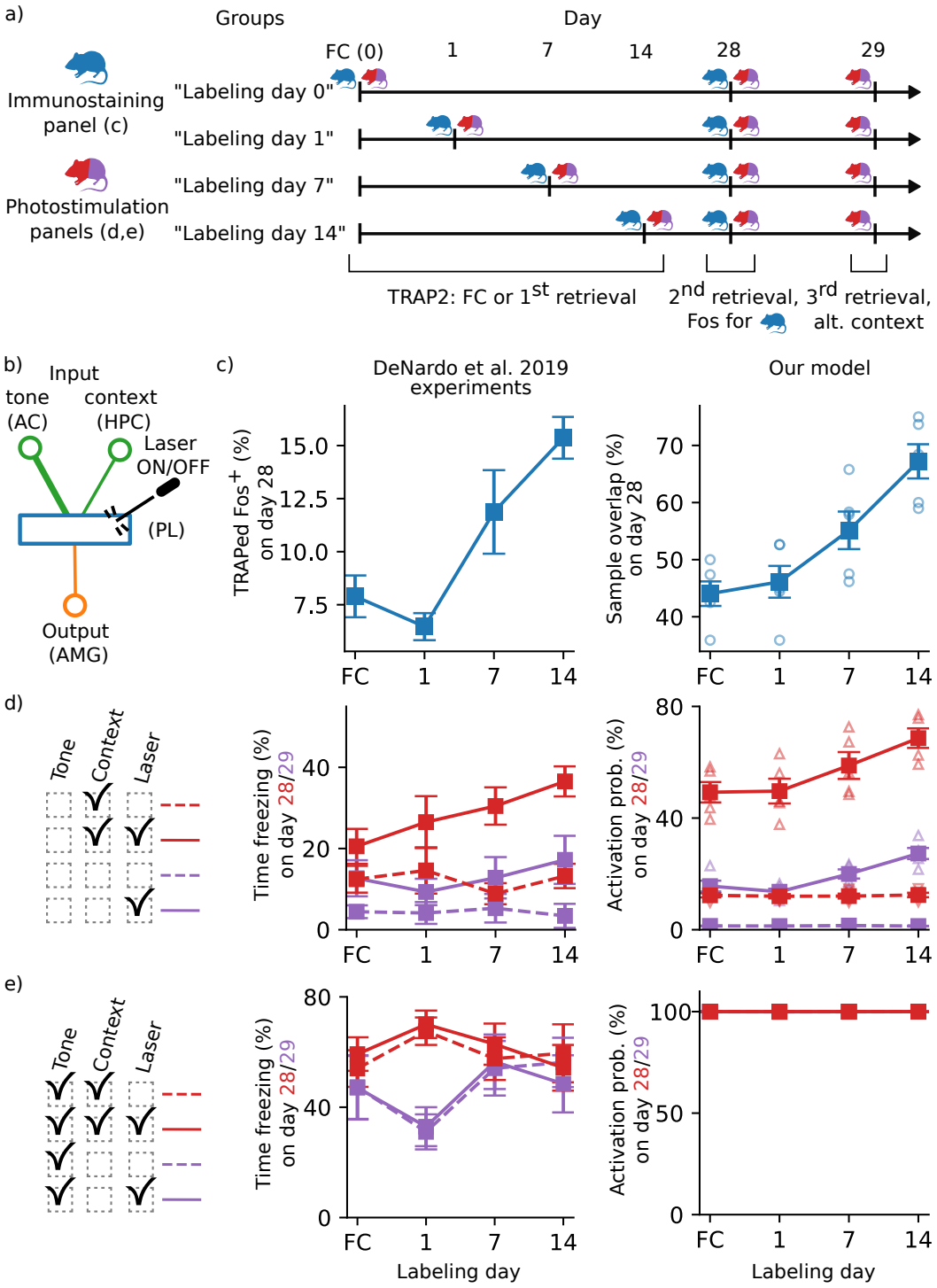

Figure 3.6: Evolution of fear memory representation observed experimentally in ref. [8] and drifting assembly model. (a) Schematics of the experiment. Immunostaining and photostimulation were conducted in different groups of mice. Immunostaining groups had two retrievals: first on day 1, 7 or 14, and second on day 28. Photostimulation groups had an additional retrieval on day 29 in an alternative context. Caption continues on next page.

Figure 3.6: (b) Model for the fear memory circuit with context input signaling conditioned cage from hippocampus (HPC), auditory stimulus input signaling conditioned tone from auditory cortex (AC), output to amygdala (AMG), and connecting drifting assembly in prelimbic cortex (PL). In some experiments a set of neurons in PL is photostimulated. Thicker line indicates stronger connection. (c, left) Experimental data [8] that are proportional to the overlap of the memory representation directly after conditioning (FC, TRAPed neurons) or during first retrieval (on day 1, 7 or 14, TRAPed neurons), with the representation at the second retrieval (day 28, Fos^+ neurons). The panel displays the fraction of Fos^+ neurons that were also TRAPed, in percent. The time interval between the labelings decreases from left to right. (c, right) Overlap of the memory assembly in our model at the beginning of the trial (FC) or on day 1, 7 or 14 (circles; squares connected by solid lines: mean), with the assembly on day 28. (d) Fear expression tested on day 28 in the conditioning context (red, context input on) and on day 29 in an alternate context (purple, context input off). Laser activating a sample of the memory representation labeled (TRAPed) on different days (as in (c)) was either on (upward triangles; squares, solid: mean) or off (downward triangles; squares, dashed: mean). Left: experiment; right: our model. (e) Same as (d) but in presence of conditioned auditory stimulus (tone input on). Red continuous lines overlay dashed and purple ones on the right. Graphs show mean ± s.e.m and data points.

this by the compensation of remodeling through the inputs and readouts. In both experiment and model, the conditioned tone input has a strong impact on behavior and results in a similar level of fear expression with and without photostimulation, in the conditioned as well as in the alternative context, Fig. 3.6e. We finally note that there is freezing in absence of any stimulus, Fig. 3.6d left. Our model replicates this since the assembly spontaneously reactivates, which also activates the output neurons.

Our results suggest that the remodeling in ref. [8] occurs due to gradual replacement of neurons in strongly connected ensembles, driven by noisy activity-dependent and spontaneous synaptic changes, which can possibly be modulated. The experiment shows only partial representation remodeling, to a level of 30% overlap between the initial and the final memory representation (Fig. 2e of [8], assuming reliable Fos staining of active neurons). The model predicts a continued decay of overlap to the chance level of the coding density, i.e. the fraction of neurons active during assembly activation (see Fig. 3.3f where it is 1/3). Specifically, this suggests that the fraction of Fos stained neurons of previously TRAPed ones (Fig. 2e of [8]) will go down to 8% (coding density, Supplementary Fig. 4d in [8]) in a several weeks longer experiment. Remodeling will continue also thereafter, Fig. 3.3f. We expect that the synaptic weights completely remodel as well, like in our model Fig. 3.3d. Finally, the model predicts that all this will happen without changing behavior. Drifting assemblies may also explain the representational change observed in [35], with partial stabilization over time resulting from reduced novelty and thus plasticity or excitability. The implementation of assemblies will likely differ, as recurrent excitation is rare in the considered brain area; the experimental predictions are nevertheless similar.

Computation with drifting assemblies

Drifting assemblies can maintain stable nonlinear computations. We have already seen this for the pattern completion computation (Appendix Figs. B.2,B.12 and accompanying text). In this paragraph we show it for the XOR gate, a classical example of a problem that is not linearly separable. Such problems cannot be solved by a simple single layer neural network (perceptron), which sums its weighted inputs and applies a threshold function. XOR has two inputs and an output. If exactly one of the inputs is active, the output gets active. Figure 3.7a shows the XOR computation realized by a circuit

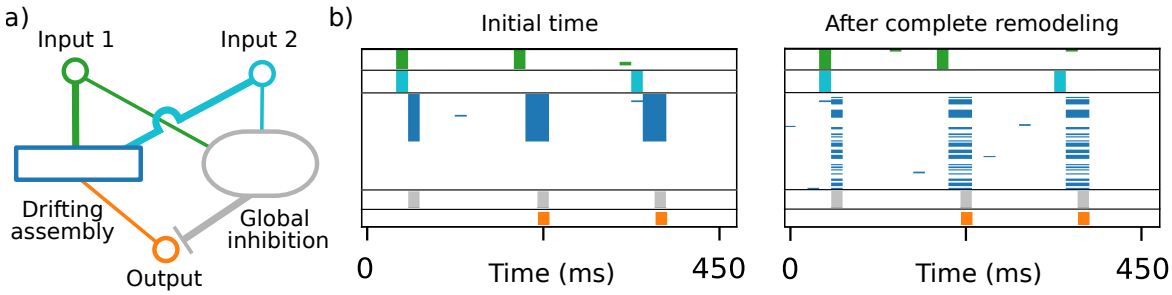

Figure 3.7: Drifting XOR gate. (a) Schematics of the network setup. Two groups of input neurons (green and cyan) are connected via a drifting assembly (blue) to output neurons (orange). An inhibitory population (gray) inhibits all neurons. Thicker lines indicate stronger connection. Only projections and neuron populations that are directly relevant for the XOR computation are shown for clarity. (b) Functionality of the drifting XOR gate at two distant times. Activity of input 1 (green), input 2 (cyan), hidden layer (blue), inhibitory (gray) and output (orange) neurons is displayed by horizontal lines in different bars (from top to bottom). The interior neurons are indexed such that initially (left) the XOR assembly activity fills the upper part of the third bar. Later, due to complete remodeling of the assembly, its activity is distributed all over the bar (right). Functionality is nevertheless conserved: the output neurons become active if exactly one of the input neuron groups is active.

of binary neurons with drifting assemblies in the hidden layer between inputs and outputs. The circuit works as follows: If a single of the two groups of XOR input neurons is activated, the XOR assembly gets active in the next step. This activates in the third step all XOR output neurons. Simultaneously it activates an inhibitory population, which suppresses the activity in the network in the step thereafter. If both sets of XOR inputs activate together, in the second step the XOR assembly activates as well. The input is, however, strong enough to also directly activate the inhibitory population. Consequently, activity in the network is suppressed in the third step. In particular the output neurons remain silent. Due to the autonomous activity in the circuit, the XOR assembly drifts and the periphery neurons follow it, which preserves the XOR input-output functionality over time, Figure 3.7b.

Autonomous network activity leads occasionally to the activation of assemblies and thereby also of the output neurons. Such events may occur only during certain brain states and therefore have no influence on behavior. Furthermore, real circuits may be composed of many assemblies, such that reactivation of one is not enough to trigger reactivation of the entire circuit. Finally, meaningful inputs will usually occur for longer periods, such that also the output stays repeatedly active over a longer period. A short, spontaneous output activation may thus be a signal that is simply ignored. Alternatively, the spontaneous reactivation may account for occasional erroneous behavior, such as spontaneous freezing (Fig. 3.6c).

Drifting assemblies in networks without spontaneous assembly reactivation

In the LIF and binary neural network models considered so far, the assemblies undergo spontaneous reactivations: events during which practically all neurons in the assembly are synchronously active. Such reactivations have been observed in various cortical areas, from visual and auditory cortex to the hippocampus [36]. To address their necessity for drifting assemblies, we use networks of linear Poisson (or “Hawkes”) model neurons [18, 37–39]. The synapses in our Poisson networks change according to a plasticity rule that does not require pre- and postsynaptic spiking, but already generates potentiation if one neuron spikes and its partner is more than on average excited [40, 41] (Methods).

Furthermore, the synapses spontaneously turn over. This drives assembly drift as in Fig. 3.5.

Assembly structure and drift are like those observed in LIF and binary network models, see Appendix Fig. B.13. The pairwise correlations in the Poisson model are, however, much smaller (Fig. 3.8a,b). They are also smaller than in the brain [42]. (The large correlations in the LIF and in binary networks are mainly due to the frequent reactivation of individual assemblies, Fig. 3.8c left and middle, Appendix Fig. B.14.)

Ignition-like reactivations as in the LIF (Appendix Figs. B.2,B.12) and binary model (Appendix Fig. B.6) can in principle not occur in networks of linear Poisson neurons, since each spike has the same impact. In contrast, the autonomous activity consists of overlapping "avalanches", transient sequences of spikes evoked by single spontaneous ones [39]. For small total synaptic weights of neurons, these are short-lived and small. An avalanche is usually confined to a part of a single assembly, since the interconnections there are still comparably strong. Avalanches involving entire assemblies do not occur on relevant time scales (Fig. 3.8c right). We observe that the occurring partial activation already suffices to sustain a drifting assembly: Its inactive neurons also have an increased level of excitation, due to the received spikes. The resulting potentiation between inactive and active neurons adds to that between the active ones. Both contributions together keep the assemblies intact and let neurons switch in conjunction with synaptic turnover.

3.3 Discussion

Our work proposes that the basis of associative memory are drifting assemblies. The drift results from fast transitions of neurons between assemblies, which are generated by noisy autonomous activity or spontaneous synaptic turnover. The represented memories nevertheless persist, because STDP and homeostasis are able to keep the representational structure intact and inputs, outputs and the assemblies consistent. This happens without requiring teaching signals or behavioral feedback, even though the network connectivity, the weight matrices and the ensembles of neurons forming the assemblies change completely. The drifting assemblies are suitable for computations and they explain recent experimental findings on changes of memory representations.

Several general suggestions have been raised on how neural network functionality may be maintained despite spontaneous synaptic turnover and unstable neural representations: First, the changes might have no impact on the relevant part of network dynamics because they are too weak or because they are eliminated by downstream attractor dynamics [1, 9, 43]. Secondly, spontaneous network remodeling may generate transitions between redundant networks that are equally well suited for the required tasks [1, 43, 44]. Thirdly, sensory feedback or feedback from other brain areas may lead to the correction of remodeling that is not in redundant directions [10, 43]. Previous computational studies addressing the impact of intrinsic synaptic fluctuations focused often on the question how neural representations can nevertheless be stable. They proposed that a preserved core structure keeps neuronal activity stable [9, 45–47], that the networks are retrained to counteract degradation [44, 48, 49] or that the ensembles of neurons storing memories are kept invariant via reactivations and unsupervised learning [21, 23, 50]. A few computational studies addressed how representations that change may emerge and partially also how the change's impact may be attenuated. Ref. [44] finds that the synaptic remodeling lets the preferred directions of hidden layer model neurons fluctuate similar to preferred directions in some recordings of motor cortex neurons. Ref. [51] shows that modest retraining of linear readouts by supervised learning allows to detect location, speed and head direction from changing

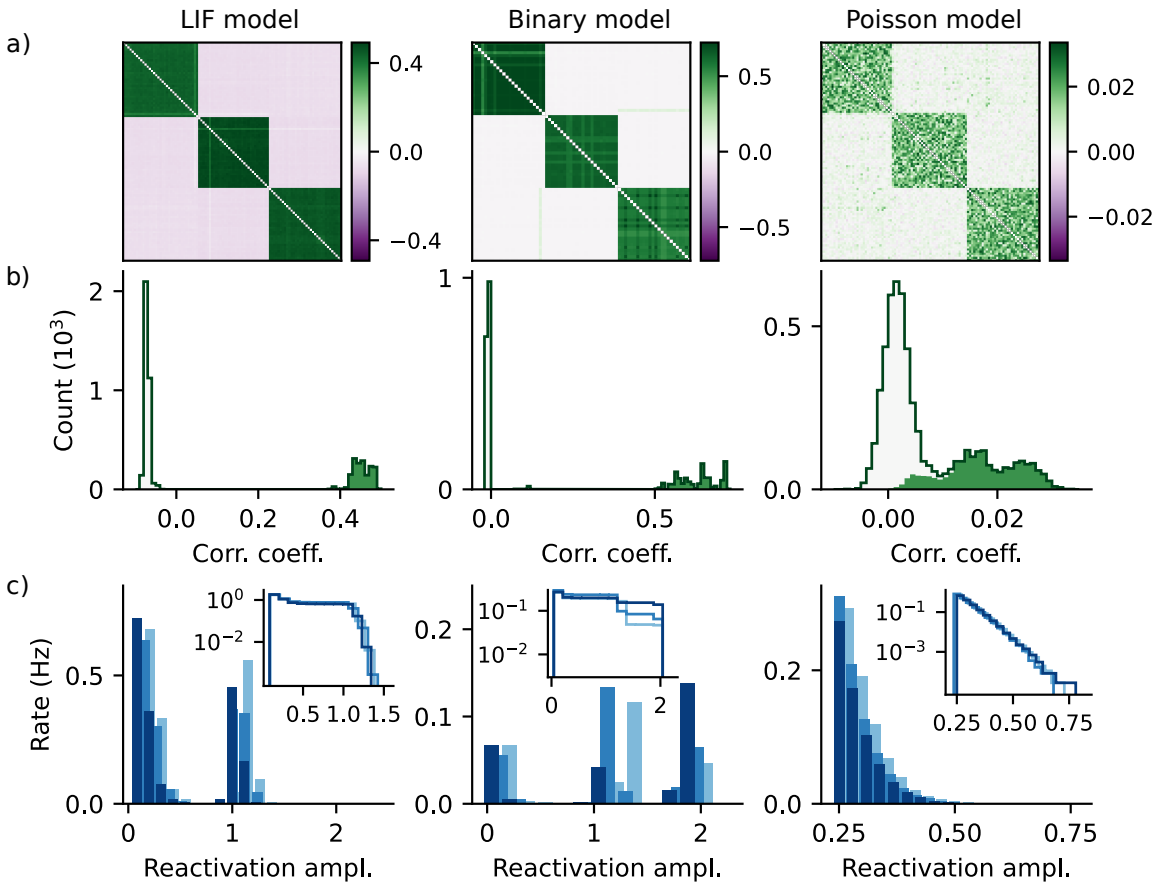

Figure 3.8: Spike correlations and assembly reactivations in our network models. Left: LIF network of Fig. 3.2. Middle: Binary network (Appendix Fig. B.6). Right: Poisson network (Appendix Fig. B.13). All networks have frozen connectivity and weights, which are obtained by fixing plastic networks after the first complete remodeling. (a),(b) Correlations between Poisson neurons (right) are much smaller than those in the other two networks (left and middle). (a) Matrices of measured spike count correlation coefficients with neurons sorted according to their assembly membership. Diagonal entries (equal to one) are blanked in white. (b) Histograms of correlation coefficients. Solid line gives the full histogram over all neuron pairs, green shading indicates the contribution from intra-assembly neuron pairs, light shading the remaining contribution from pairs of neurons belonging to different assemblies. (c) Assembly reactivation is absent in Poisson networks. Main panels show how often different maxima of the number of spikes in a moving time window are detected within an assembly (Methods). Occurrence rate is plotted against number of spikes divided by the assembly size (different blues for the assemblies). Events of spontaneous assembly reactivation are reflected by a second peak, separate from the background continuum near zero. Insets show the complementary cumulative distributions with logarithmic rate scale.

representations in the posterior parietal cortex. Ref. [52] found in a model for developing zebrafish networks spontaneously emerging and changing assemblies, which break down and merge and thus seem unsuitable for persistent representation. Ref. [53] observed merging of assemblies. Ref. [47] shows that if a fluctuating part is added to otherwise stable synaptic weights storing a sequence, individual neurons partially and temporarily change their dynamics, but the original sequence can still be detected.

In our networks there is no preserved core of network structure or memory representation. Only the full circuit consisting of inputs, assemblies and outputs preserves memory and behavior. Interior neurons switch from coding for one task to coding for another, only periphery neurons have stable codes. The inputs and the readouts therefore have to constantly compensate for remodeling, by unsupervised plasticity combining STDP and homeostasis. This requires no external feedback and no error signal exchange within the brain, as collective activity tells the neurons where to attach. Such unsupervised, self-organized compensation may occur for other types of drifting representations as well. Further it may compensate other perturbations like moderate neuronal death. We note that a corrective feedback from other brain areas would require a stable anchoring in an invariant representation, otherwise there will be an error in the error signal, thus co-evolution of errors and finally erroneous behavior. On the level of synapses there is constant change in both periphery and interior neurons in our model; no synapse is persistent. This predicts that in memory areas weight remodeling will turn out to be complete [7].

The developed models combine established model neurons [12, 14, 37], simple STDP or correlation learning [12, 18, 26] and homeostasis [18, 28, 29, 52]. Robustness is indicated by the occurrence of drifting assemblies in different neuron models, the simple mechanisms underlying the drift and the broad range of correlation strengths that can sustain drifting assemblies; the range includes the correlations found in biological neural networks [42]. The number of assemblies in the investigated networks is largely fixed by the network parameters. We also observe a dependence on the initialization (Appendix Fig. B.5), which might lead to more variability for larger networks with more assemblies. The approximately constant number of assemblies renders our memory networks less flexible. Versatile learning could still be implemented by input-dependent imprinting of new assemblies. This would lead to the vanishing or merging of existing ones and, if these already represent memories, to forgetting or generalization. It is not known whether the number of assemblies changes in the brain; it might be that assemblies form during development [54]. Preexisting assemblies could become meaningful when specific inputs and outputs are assigned to them through learning, similar to preexisting sequential structures in a model of episodic memory [55]. Our models can generate very different drifting times from hours to years, which depend on underlying parameters like noise strength and synaptic lifetime. Furthermore, the drifting of assemblies is consistent with the requirement of executing nonlinear computations with them. Finally, assuming that drifting assemblies form the memory representations in the prelimbic cortex of mice explains the main and several detailed experimental findings in ref. [8]. We therefore propose that assembly drift as predicted by our models occurs in biological neural networks.

We find that noisy autonomous activity and the resulting synaptic weight changes can cause transitions of neurons between assemblies. This is due to noise-driven switching between meta-stable states and inhomogeneous noise that prevents neurons to linger between assemblies. Further our findings suggest a specific connection between spontaneous synaptic change and the remodeling of neural representations: spontaneous synaptic changes push individual neurons out of their assembly and lets them switch to new ones. Since in biological neural networks spontaneous synaptic change is

significant [3, 5, 7] and activity is noisy [12, 42], we suggest that both together drive the drifting of assemblies in biological neural networks. The strength of input synapses thereby decides whether a neuron continues to stay with its assembly, suggesting that input neurons rely on feedback from their assembly to be able to follow it. Our model thus provides an additional explanation for the abundance of feedback connections throughout cortical areas, including early sensory ones.

Our model suggests that the changes through assembly drift are functionally per se neutral. Modification and acquisition of memories will happen on top. The drift might nevertheless help the storage of memories. First, it may contribute to solving the stability-plasticity dilemma: New memories can be imprinted in the same highly plastic (sub-)region (cf. also the two-stage model for memory [11]). With time their representations drift away together with their input and output connections, giving way to further ones and joining older representations in less plastic regions. Second, in our networks, inputs, outputs and assemblies need to stay consistent for memory storage. Having static assemblies is an additional requirement. It may in the brain be difficult to realize due to noisy activity, ongoing plasticity and spontaneous synaptic change. The proposed self-organized compensation may have abolished the evolutionary pressure to develop static assemblies; the synaptic plasticity rules were less constrained and free to satisfy other requirements.

Our results make specific and general predictions for memory systems. In particular, they predict a continued drift, reduction of overlap to chance level and gradual but complete remodeling of synaptic weights, as well as the mechanisms underlying the drift and ways to interfere with it. On a more general level, our results indicate that impairments of synaptic plasticity do not only lead to problems in acquiring new memories but also to forgetting, either because assemblies break down or because the improperly compensating inputs and outputs disconnect from their drifting assemblies. Changes in synaptic plasticity as observed in dementia may thus directly lead to the symptomatic forgetting. Further, unsupervised compensation of drifting representations could be a common theme in the brain. Finally, our work suggests that Heraclitus of Ephesus' 2500 year old idea is also true for memory representation, namely that: πάντα ῥεῖ; everything flows or, in other words: drifts.

3.4 Materials and methods

To demonstrate the feasibility of drifting assemblies we consider networks of LIF model neurons. For more specific questions we use further simplified models of Poisson spiking [18, 37–39] or binary neurons [14, 28, 29, 52]. The long-term simulations require that we consider networks of medium size, of the order of a hundred neurons, storing between two and ten memories (Appendix Fig. B.11). The simplest networks where we can observe neuron switching between assemblies store two assemblies. In networks with three and more assemblies, however, switching includes an additional feature, choosing an assembly to switch to. We thus use networks with three assemblies for our proof of principle simulations. Since the memories are in our model stored in the synapses between excitatory neurons, only these are plastic. The remaining synapses are homogeneous (LIF, binary model) or not explicitly modeled (Poisson model), such that there is no static network weight structure that could contribute to the maintenance of network functionality.

Previous theoretical work has shown that learning [16, 17] or spontaneous emergence [18, 22] of static assemblies as well as their maintenance can be enabled by a combination of STDP and homeostatic plasticity. A recent experimental study investigated STDP in the recurrent excitatory synapses of the hippocampal region CA3 [26], i.e. of a region that is assumed to serve as an associative

memory network and store assemblies [11]. It found an STDP learning window with LTP for pre- and post-synaptic spikes, irrespective of their ordering and the stronger the closer they are. Based on the theoretical and experimental results we choose in our models an STDP rule with a symmetric learning window with centrally peaked LTP (Appendix Fig. B.1). LTP is compensated by LTD for temporally distant spikes, see also [18]. In networks of binary neurons we reduce the STDP to a standard Hebbian covariance rule [12, 52] and augment it by a dependence of weight changes on previous weight. Finally, in our networks of Poisson neurons we use an STDP rule that is similar to voltage- or calcium-based rules [40, 41]. In all models the synapses further undergo homeostasis. Experiments indicate that the total input [28, 56, 57] to a neuron may be conserved. Further, there is evidence for output normalization [58]. Both input and output normalization may be realized by competition for synaptic resources [28]. Following [18, 28, 29, 52], we thus introduce normalizations of the model neuron's input and output weights. Finally, we incorporate spontaneous synaptic turnover in some of the models.

LIF networks

We use a current-based LIF neuron model. The membrane potential $V_i(t)$ of neuron i obeys

$$\tau_m \frac{dV_i(t)}{dt} = V_{\text{rest}} - V_i(t) + RI_i^E(t) + RI_i^I(t) + \sqrt{2\tau_m}\sigma\xi_i(t), \tag{3.1}$$

$$RI_i^E(t) = \sum_{j\in M_E} w_{ij} \sum_{t_j\le t} e^{-\frac{t-t_j}{\tau_E}},\ RI_i^I(t) = \sum_{j\in M_I} w_{ij} \sum_{t_j\le t} e^{-\frac{t-t_j}{\tau_I}}. \tag{3.2}$$

Here, τ_m is the membrane time constant, V_{rest} the resting membrane potential, R the input resistance, $I_i^E(t)$ and $I_i^I(t)$ are the total currents generated by the populations of excitatory and inhibitory synapses and $\xi_i(t)$ is standard Gaussian white noise. The parameter σ equals the standard deviation of the membrane potential's stationary distribution in absence of a threshold and synaptic input currents from the other modeled neurons; the membrane potential then follows an Ornstein-Uhlenbeck process. When the voltage exceeds a spike threshold $V_\theta = 20\text{mV}$, a spike is generated and the neuron is reset to $V_0 = 0\text{mV}$, where it stays for a refractory period τ_{ref}. V_{rest} is halfway between threshold and reset. A generated spike travels to postsynaptic neurons, where it generates changes in the synaptic currents. A spike of an excitatory neuron j evokes in the input $RI_i^E(t)$ a jump-like increase of height w_{ij}, which thereafter decays exponentially with time constant τ_E. The decay time constant of inhibitory input currents is τ_I. t_j are the spike times of neuron j, M_E is the set of all N_E excitatory and M_I that of all N_I inhibitory neurons.

At each excitatory spike, the synaptic weights are updated according to a pair-based spike rule with symmetric STDP window (Appendix Fig. B.1). Each side of the window is the difference of two exponentials with decay time constants $\tau_{\text{LTP}} = 20\text{ms}$ and $\tau_{\text{LTD}} = 40\text{ms}$. In terms of the induced change of peak EPSP, peak LTP is 0.50mV (0.17mV), at 0ms, peak LTD is −0.17mV (−0.06mV), at ±44ms, for connections between interior neurons in the network of Fig. 3.2 (connections from or to periphery neurons in the network of Fig. 3.2 and all connections in the network of Fig. 3.5, i.e. STDP is weaker in the network with spontaneous synaptic turnover). Homeostatic plasticity normalizes the total excitatory input and output weight strength of an interior neuron i to $\sum_{j\in M_E} w_{ij} = \sum_{j\in M_E} w_{ji} = w_{\text{sum}}$; for a periphery neuron the total input and output weight strength is $w_{\text{sum,peri}}$. The normalization is approximated by divisively normalizing after each excitatory spike the columns and the rows of the

weight matrix between the excitatory neurons. The excitatory weights between interior neurons are bounded by 0mV and $w_{\max}$, for weights from or to periphery neurons the upper bound is $w_{\max,\text{peri}}$. These bounds are enforced by clipping weights before and after homeostatic normalization. All possible synapses between excitatory and inhibitory and between inhibitory neurons are present. There are no self-connections.

Assemblies

The memories are encoded in binary, non-graded manner [59]. In particular (almost) all neurons are strongly activated, if their assembly is active. Persistent memory manifests itself as a conserved behavioral input-output relation: the input neurons that activate the different ensembles forming an assembly are not exchanged over time as well as the output neurons activated by them. For simplicity, we assume that memories and the related behaviors are unchanged over time. We further assume that each input and output neuron is specific to one representation and we setup the assemblies without shared neurons. Networks are initialized by setting existing weights between neurons within an assembly and between an assembly and its periphery neurons to 1mV and all others to 0mV, then homeostatic normalization and clipping are applied.

Spontaneous synaptic turnover

A synapse between two excitatory neurons in our networks with spontaneous synaptic turnover has a finite expected lifetime L. It vanishes with rate $1/L$, i.e. in a simulation step of duration Δt with probability $\Delta t/L$, independent of activity and of its current weight. Similarly, if the synapse is absent, it has an average absence time A; it appears in a simulation step with probability $\Delta t/A$. Thus, on average the synapse is present a fraction $L/(L+A)$ of the time; the probability that it is present at a certain time point is $p = L/(L+A)$, the density of synapses of a certain type. The spontaneous turnover in our networks switches entries of the connectivity matrix between 1 and 0. When a synapse vanishes, its weight becomes 0mV. Newly appearing synapses have weight 0mV as well [6].

Simulation of LIF networks and analysis

The networks in Figs. 3.2 and 3.5 consist of $N_E = 102$ excitatory and $N_I = 20$ inhibitory neurons. $N_{\text{int}} = 90$ of the excitatory neurons are interior neurons. These are initiated such that there are three assemblies of $N_{\text{asbly}}(0) = 30$ interior neurons. Each assembly has 4 periphery neurons. The first two periphery neurons are designated input, the second two output neurons. Connectivity in the network of Fig. 3.2 is all-to-all; the connection density is high to compensate small network size. In Fig. 3.5 the connection density is $p_{\text{int}} = 0.6$ between interior neurons and $p_{\text{peri}} = 0.8$ between interior and periphery neurons. The corresponding synaptic life and absence times are for synapses between interior neurons $L_{\text{int}} = 2000$s and $A_{\text{int}} = 1333.3$s and for those between interior and periphery neurons $L_{\text{peri}} = 2000$s and $A_{\text{peri}} = 500$s. Four periphery neuron weights exceed the range of the colorbar in Fig. 3.5; the largest weight would evoke a 3.9mV high EPSP. We did five alike simulations each with total simulation time 75 hours (100 hours for the network in Fig. 3.5) and different random realizations of networks and noise to check that the representational structure is conserved over time, i.e. that assemblies continuously drift and their periphery neurons faithfully follow them. Clustering is throughout the article obtained with the Louvain clustering algorithm [60] as implemented in [61].

Fig. 3.3a shows the sum of the weights between interior neuron 2 (indexed 6 in Fig. 3.2d) and assemblies 1,2 and 3, normalized by $2w_{\text{sum}}$ (total input plus total output weight). (b) similarly shows the sum of the weights between input neuron 1 and assemblies 1,2 and 3 normalized by $2w_{\text{sum,peri}}$. (e) The Pearson correlation between weight matrices at times 0 and t is computed as $\text{Corr}(t) = \left(\sum_{ij} \tilde{w}_{ij}(0)\tilde{w}_{ij}(t)\right) / \left(\sqrt{\sum_{ij} \tilde{w}_{ij}(0)^2}\sqrt{\sum_{ij} \tilde{w}_{ij}(t)^2}\right)$, where $\tilde{w}_{ij} = w_{ij} - \bar{w}$ are the matrix entries centered by the average entry size $\bar{w}$. The maximal correlation is 1. (f) Displayed is the sum of the weights between the neurons originally forming the first, second, third assembly (dark shadings of blue), normalized by their maximal sum $N_{\text{asbly}}(0)w_{\text{sum}}$. Further displayed is the sum of weights between the neurons originally forming the first and second, first and third, second and third assembly (light shadings of blue), normalized by their maximal sum $2N_{\text{asbly}}(0)w_{\text{sum}}$. The inset shows the analogous quantities for the current assemblies at each time, normalized using the current assembly sizes. Chance level at a time is computed by summing all weights between interior neurons and normalizing by $N_{\text{int}}w_{\text{sum}}$. Panel (g) shows the overlap of the realizations of assembly 1 with previous and future reference ensembles. We compute the overlap as the number of neurons that an ensemble of neurons shares with its reference ensemble, normalized by the size of the reference ensemble. The overlap is thus bounded by 0 and 1. We use as criterion for complete remodeling of an assembly with respect to a previous reference ensemble that the overlap with the reference has decreased to chance level.

We use as criterion for complete network remodeling that for all assemblies the overlap with their original realization has reached chance level at least once.

Statistics of weight changes

In Fig. 3.4 we measure average weight changes between neurons and assemblies and their fluctuations. We record for an interior neuron the change Δw_1 of its summed input weight w_1 from assembly 1, in successive time intervals whose lengths equal the average single neuron interspike interval. We repeat the process for all interior neurons and their inputs from all assemblies. The weights are normalized by w_{sum}, such that w_1 is between 0 and 1. We bin this range into 50 bins of size 0.02 and calculate for each bin the average $\overline{\Delta w_1}$ and standard deviation $\text{Std}(\Delta w_1)$ of the changes ensuing those w_1 that fall in it. To obtain the potential $U(w_1)$, we think of the average $\overline{\Delta w_1}(w_1)$ as being evoked by a force, $F(w_1) = \overline{\Delta w_1}(w_1)$, like in a classical mechanics system where friction dominates over negligible inertia. The potential $U(w_1)$ determines the force by $F(w_1) = -dU(w_1)/dw_1$ (gradient system). $U(w_1)$ is computed by integrating $-\overline{\Delta w_1}(w_1)$ over w_1.

Binary model

The dynamics of the binary model are given by

$$x_i(t) = \Theta\left(\sum_j w_{ij}x_j(t-1) + I^I(t-1) - \theta\right), \tag{3.3}$$

where $x_i(t) \in \{0, 1\}$ is the activity of the excitatory neuron i, $i \in \{1, ..., N\}$, at time step t, w_{ij} the weight matrix of excitatory-to-excitatory connections, θ a neuron firing threshold and Θ the Heaviside step function. If present, periphery neurons are not interconnected. There are no self-connections.

$I^I(t-1)$ is the inhibitory input. We use global inhibition that depends on the average excitatory activity, $I^I(t) = -\Theta\left(1/N \sum_j x_j(t) - \theta_I\right)$, where θ_I is an inhibitory unit firing threshold. At every time step each neuron is stimulated with probability p_{sp} to spike spontaneously.

The Hebbian learning rule, applied in every step, has the form $\Delta w_{ij}(t) = \eta(w_{ij})(x_i(t) - \mu_i)(x_j(t) - \mu_j)$, with a long-term average μ_i of the activity of neuron i (covariance rule), and learning rate $\eta(w_{ij})$. $\eta_{ij}(w_{ij}) = \eta_{\text{weak}}$ if $w_{ij} < w_{\text{th}}$ and η_{strong} otherwise. The values of η_{weak} and w_{th} differ between interior and periphery neurons, Appendix Fig. B.1. All weights w_{ij} are bounded between 0 and the same $w_{\max}$. After updating the weights with the learning rule, first the outputs and then the inputs of all neurons are normalized to $w_{\text{sum}} = 1$.

Model for fear memory representation

The model of fear memory representations is implemented with 150 excitatory binary neurons. The 117 interior neurons initially realize three assemblies of equal size. Each assembly has 11 periphery neurons. One assembly is chosen to represent the fear memory. Its periphery neurons are split into two context input neurons, six tone input neurons and three output neurons. At fear conditioning and each of the first recalls 100 samples of 4 interior neurons of the current fear memory realization are selected. These will be activated by the photostimulation during testing. The final recall is modeled as detection of the fear memory assembly on day 28. Thereafter the overlaps with previous assembly realizations are computed (cf. Fig. 3.6b) and plasticity is turned off to test the response of the system under different conditions (Fig. 3.6c,d). Each test lasts for 5 time steps (75ms), during the first three steps the appropriate input and photostimulated neurons are active. The circuit is considered to produce a relevant output signal, if all output neurons spike together in at least one of the last four time steps of the test. For each experimental condition, we compute the probability of such an output by averaging over 50 testing runs, for each sample of photostimulated neurons and these samples. We simulate and evaluate overall five different realizations of the system to emulate experiments with five different animals. Fig. 3.6 shows the overlaps and output activation probabilities for each of these realizations, together with the means taken over realizations and their standard errors.

XOR gate

The XOR gate model is setup as follows: The gate inputs and outputs are represented by groups of periphery neurons. The hidden layer consists of excitatory interior neurons and an inhibitory population. The interior neurons form two drifting assemblies. One of them connects to the XOR gate periphery neurons, the other is unrelated to the computation. The inhibitory population receives input from the periphery and the interior neurons; it inhibits all neurons in the network. Specifically, our system consists of 100 binary neurons, split into two assemblies of 36 neurons with 14 periphery neurons each. All inhibitory populations are modeled by single units. The XOR assembly has two input ensembles, each with 6 neurons, and 2 output neurons. Two additional inhibitory populations are added to the XOR input neurons. These receive excitation from one of the two input ensembles and get activated if all their inputs spike. If they are active, they inhibit both input ensembles. This additional inhibition is introduced to prevent the the activation of the inputs by the assembly, which would lead to both inputs being activated when the output is activated. Time steps are 15ms long.

Linear Poisson model

In linear Poisson neurons each spike has the same impact, independent of the current state of the neuron. This fits to background activity, where the state of a neuron stays close to a baseline all the time. Linear Poisson neurons are stochastically spiking neurons with instantaneous rates $f_i(t)$, $i = 1, \ldots, N$, evolving in continuous time. The rates are excited by spikes from the other neurons in the network and follow the linear dynamics

$$\tau \frac{d}{dt} f_i(t) = f_0 - f_i(t) + \tau \sum_{j=1}^{N} w_{ij} \sum_{t_j} \delta\left(t - t_j\right), \tag{3.4}$$

where f_0 is a constant spontaneous rate due to the assumed embedding in a fluctuation-driven asynchronous irregular activity state. A spike from neuron j increases $f_i(t)$ in a jump-like manner by a nonnegative synaptic weight w_{ij}. Between input spikes $f_i(t)$ decays exponentially with time constant $\tau = 10$ms to the spontaneous rate $f_0 = 0.75$Hz. On average a spike of neuron j induces τw_{ij} additional spikes in neuron i. Global or explicitly modeled inhibition is assumed to be implicitly contained in the model; it contributes to the randomness of spike generation, both of spontaneous and of excited activity. The network dynamics can be solved in an event-based manner allowing for very long simulation times.

The synapses in our Poisson networks change according to the following plasticity rule. When a neuron spikes, its existing input and output synapses are changed depending on the current level of excitation of the corresponding partner neurons, which we measure by the instantaneous rate above baseline, $f_i(t) - f_0$. The dependence is given by a function $\Delta w(f_i - f_0)$, which is negative (giving rise to LTD) for small and average values of the excitation and positive (giving rise to LTP) for larger values; for simplicity we use a quadratic function (Appendix Fig. B.1) with $\Delta w(0) = 0$. The weights are bounded by 0 and $w_{\max}$. We model homeostatic plasticity by input and output normalization of summed synaptic weights as in the other network models. Here this implies that the average number of additional spikes in the network induced by a spike of neuron j (also referred to as 'branching parameter') remains constant: $\tau \sum_i w_{ij} = \tau w_{\text{sum}} = 0.25$. The synaptic connections in our Poisson networks also turn over spontaneously as described above.

Correlations and reactivation amplitudes

To quantify the pairwise spike correlations in our network models in Fig. 3.8a,b, we use the Pearson correlation coefficients of spike counts [42] in time bins of size 150 ms. We measure the spike counts in simulations of static networks with connectivity and weights taken from simulations of networks with drifting assemblies, after the first complete remodeling. The measurement time is 2 h for the LIF, 15 h for the binary and 10 h for the Poisson network. The neurons of the static networks are sorted according to their detected assembly membership as in the figures demonstrating drifting assemblies. This allows us to partition the neuron pairs into intra-assembly pairs (where both neurons are in the same assembly) and inter-assembly pairs (where each neuron is in a different assembly) and to show their contributions to the distributions of correlation coefficients in the networks (Fig. 3.8b).

To investigate assembly reactivation, in Fig. 3.8c we consider the summed spiking activity of all neurons in an assembly. This assembly spiking activity is then temporally filtered with a moving time window of size 15 ms for the LIF, 45 ms for the binary and 100 ms for the Poisson model. Reactivation

events in our LIF and binary networks are equal to or shorter than the corresponding window size. Further, the analytical duration distribution [39] shows that for our Poisson networks' branching parameter and time constant only 0.013% of single avalanches are longer than 100 ms, justifying the latter window size. The filtering results in a time series, which gives at each time the number of spikes that have occurred in an assembly during the preceding time window. We locate local maxima of this time series to detect putative assembly reactivation. If two local maxima are found with a temporal distance less than the filter window size, only the larger one is kept. The heights of the local maxima are initially numbers of spikes; we normalize them by the corresponding assembly size and call them (relative) reactivation amplitudes. We only consider amplitudes above or equal to a minimal value (0.125 for LIF, 0.125 for binary and 0.25 for Poisson networks). Fig. 3.8c shows histograms of the amplitudes obtained from the same simulations as the correlation measurements. Dividing the counts of different amplitudes by the measurement time gives the displayed occurrence rate. The associated complementary cumulative distributions (Fig. 3.8c, insets) indicate the rates at which putative assembly reactivation with sizes exceeding the given amplitude occur.

3.5 Acknowledgments

A thank Paul Züge for fruitful discussions, Abigail Morrison for comments on the manuscript, Hans Günter Memmesheimer, Katharina Hack and Jonas Nietzsche for help with the graphical illustrations, and the German Federal Ministry of Education and Research (BMBF) for support via the Bernstein Network (Bernstein Award 2014, 01GQ1710).

References

[1] S. Rumpel and J. Triesch, *The dynamic connectome*, e-Neuroforum **22** (2016).

[2] Y. Humeau and D. Choquet, *The next generation of approaches to investigate the link between synaptic plasticity and learning*, Nature Neuroscience **22** (2019) 1536.

[3] N. Yasumatsu, M. Matsuzaki, T. Miyazaki, J. Noguchi and H. Kasai, *Principles of Long-Term Dynamics of Dendritic Spines*, Journal of Neuroscience **28** (2008) 13592.

[4] A. Rubinski and N. E. Ziv, *Remodeling and Tenacity of Inhibitory Synapses: Relationships with Network Activity and Neighboring Excitatory Synapses*, PLoS Computational Biology **11** (2015) e1004632, ed. by K. T. Blackwell.

[5] R. Dvorkin and N. E. Ziv, *Relative Contributions of Specific Activity Histories and Spontaneous Processes to Size Remodeling of Glutamatergic Synapses*, PLoS Biology **14** (2016) e1002572, ed. by E. M. Schuman.

[6] K. P. Berry and E. Nedivi, *Spine Dynamics: Are They All the Same?*, Neuron **96** (2017) 43.

[7] N. E. Ziv and N. Brenner, *Synaptic Tenacity or Lack Thereof: Spontaneous Remodeling of Synapses*, Trends in Neurosciences **41** (2018) 89.

[8] L. A. DeNardo et al.,
Temporal evolution of cortical ensembles promoting remote memory retrieval,
Nature Neuroscience **22** (2019) 460.

[9] C. Clopath, T. Bonhoeffer, M. Hübener and T. Rose,
Variance and invariance of neuronal long-term representations,
Philosophical Transactions of the Royal Society B: Biological Sciences **372** (2017) 20160161.

[10] M. E. Rule, T. O'Leary and C. D. Harvey, *Causes and consequences of representational drift*,
Current Opinion in Neurobiology **58** (2019) 141.

[11] G. Buzsáki, *Neural Syntax: Cell Assemblies, Synapsembles, and Readers*,
Neuron **68** (2010) 362.

[12] W. Gerstner, W. M. Kistler, R. Naud and L. Paninski,
Neuronal Dynamics - From single neurons to networks and models of cognition,
Cambridge: Cambridge University Press, 2014.

[13] G. Mongillo, S. Rumpel and Y. Loewenstein, *Intrinsic volatility of synaptic connections — a challenge to the synaptic trace theory of memory*,
Current Opinion in Neurobiology **46** (2017) 7.

[14] A. Scott, *Neuroscience: a Mathematical Primer*, Springer New York, 2002,
URL: https://doi.org/10.1007/b98897.

[15] T. P. Vogels, H. Sprekeler, F. Zenke, C. Clopath and W. Gerstner, *Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks.*, eng,
Science **334** (2011) 1569.

[16] A. Litwin-Kumar and B. Doiron,
Formation and maintenance of neuronal assemblies through synaptic plasticity., eng,
Nature Communications **5** (2014) 5319.

[17] F. Zenke, E. J. Agnes and W. Gerstner, *Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks.*, eng,
Nature Communications **6** (2015) 6922.

[18] N. Ravid Tannenbaum and Y. Burak,
Shaping Neural Circuits by High Order Synaptic Interactions,
PLoS Computational Biology **12** (2016) 1.

[19] G. K. Ocker and B. Doiron,
Training and Spontaneous Reinforcement of Neuronal Assemblies by Spike Timing Plasticity,
Cerebral Cortex **29** (2018) 937.

[20] J. Herpich and C. Tetzlaff,
Principles underlying the input-dependent formation and organization of memories,
Network Neuroscience **3** (2019) 606.

[21] J. Humble, K. Hiratsuka, H. Kasai and T. Toyoizumi, *Intrinsic Spine Dynamics Are Critical for Recurrent Network Learning in Models With and Without Autism Spectrum Disorder*,
Frontiers in Computational Neuroscience **13** (2019).

[22] L. Montangie, C. Miehl and J. Gjorgjieva,
Autonomous emergence of connectivity assemblies via spike triplet interactions,
PLoS Computational Biology **16** (2020) 1.

[23] M. J. Fauth and M. C. van Rossum,
Self-organized reactivation maintains and reinforces memories despite synaptic turnover,
eLife **8** (2019).

[24] L. Wittgenstein, *Philosophische Untersuchungen/Philosophical investigations*,
ed. by P. M. S. Hacker and J. Schulte, Oxford: Wiley-Blackwell, 2009.

[25] T. Hainmueller and M. Bartos,
Parallel emergence of stable and dynamic memory engrams in the hippocampus,
Nature **558** (2018) 292.

[26] R. K. Mishra, S. Kim, S. J. Guzman and P. Jonas, *Symmetric spike timing-dependent plasticity at CA3–CA3 synapses optimizes storage and recall in autoassociative networks*,
Nature Communications **7** (2016).

[27] E. Bienenstock, L. Cooper and P. Munro, *Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex*,
Journal of Neuroscience **2** (1982) 32.

[28] I. R. Fiete, W. Senn, C. Z. H. Wang and R. H. R. Hahnloser,
Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity., eng, Neuron **65** (2010) 563.

[29] A. Lazar, G. Pipa and J. Triesch, *SORN: a self-organizing recurrent neural network.*, eng,
Frontiers in Computational Neuroscience **3** (2009) 23.

[30] R. Q. Quiroga, *Plugging in to Human Memory: Advantages, Challenges, and Insights from Human Single-Neuron Recordings*, Cell **179** (2019) 1015.

[31] C. Gardiner, *Handbook of Stochastic Methods*, Berlin: Springer, 2002.

[32] W. Horsthemke and R. Lefever, *Noise-Induced Transitions*, Berlin: Springer, 1984.

[33] D. B. Headley, V. Kanta, P. Kyriazi and D. Paré, *Embracing Complexity in Defensive Networks*,
Neuron **103** (2019) 189.

[34] N. Karalis et al.,
4-Hz oscillations synchronize prefrontal–amygdala circuits during fear behavior,
Nature Neuroscience **19** (2016) 605.

[35] A. Attardo et al.,
Long-Term Consolidation of Ensemble Neural Plasticity Patterns in Hippocampal Area CA1,
Cell Reports **25** (2018) 640.

[36] A. U. Sugden et al., *Cortical reactivations of recent sensory experiences predict bidirectional network changes during learning*, Nature Neuroscience **23** (2020) 981.

[37] R. Kempter, W. Gerstner and J. L. Van Hemmen, *Hebbian learning and spiking neurons*,
Physical Review E **59** (1999) 4498.

[38] V. Pernice, B. Staude, S. Cardanobile and S. Rotter,
How structure determines correlations in neuronal networks,
PLoS Computational Biology **7** (2011) e1002059.

[39] Y. F. Kalle Kossio, S. Goedeke, B. van den Akker, B. Ibarz and R.-M. Memmesheimer,
Growing Critical: Self-Organized Criticality in a Developing Neural System,
Physical Review Letters **121** (5 2018) 058301.

[40] D. V. Buonomano,
A Learning Rule for the Emergence of Stable Dynamics and Timing in Recurrent Networks,
Journal of Neurophysiology **94** (2005) 2275.

[41] C. Clopath, L. Büsing, E. Vasilaki and W. Gerstner,
Connectivity reflects coding: a model of voltage-based STDP with homeostasis,
Nature Neuroscience **13** (2010) 344.

[42] A. Renart et al., *The Asynchronous State in Cortical Circuits*, Science **327** (2010) 587.

[43] A. R. Chambers and S. Rumpel, *A stable brain from unstable components: Emerging concepts and implications for neural computation*, Neuroscience **357** (2017) 172.

[44] U. Rokni, A. G. Richardson, E. Bizzi and H. S. Seung,
Motor Learning with Unstable Neural Representations, Neuron **54** (2007) 653.

[45] G. Mongillo, S. Rumpel and Y. Loewenstein,
Inhibitory connectivity defines the realm of excitatory plasticity,
Nature Neuroscience **21** (2018) 1463.

[46] L. Susman, N. Brenner and O. Barak, *Stable memory with unstable synapses*,
Nature Communications **10** (2019).

[47] M. Gillett, U. Pereira and N. Brunel,
Characteristics of sequential activity in networks with temporally asymmetric Hebbian learning,
Proceedings of the National Academy of Sciences **117** (2020) 29948.

[48] R. Ajemian, A. D'Ausilio, H. Moorman and E. Bizzi, *A theory for how sensorimotor skills are learned and retained in noisy and nonstationary neural circuits*,
Proceedings of the National Academy of Sciences **110** (2013) E5078.

[49] D. Kappel, R. Legenstein, S. Habenschuss, M. Hsieh and W. Maass,
A Dynamic Connectome Supports the Emergence of Stable Computational Function of Neural Circuits through Reward-Based Learning, eNeuro **5** (2018) 0301.

[50] D. Acker, S. Paradis and P. Miller,
Stable memory and computation in randomly rewiring neural networks,
Journal of Neurophysiology **122** (2019) 66.

[51] M. E. Rule et al., *Stable task information from an unstable neural population*,
eLife **9** (2020) e51121, ed. by S. Palmer and R. L. Calabrese.

[52] M. A. Triplett, L. Avitan and G. J. Goodhill, *Emergence of spontaneous assembly activity in developing neural networks without afferent input*, PLoS Computational Biology **14** (2018) 1.

[53] N. Hiratani and T. Fukai,
Interplay between Short- and Long-Term Plasticity in Cell-Assembly Formation,
PLoS One **9** (2014) 1.

[54] T. Pietri et al., *The Emergence of the Spatial Structure of Tectal Spontaneous Activity Is Independent of Visual Inputs*, Cell Reports **19** (2017) 939.

[55] S. Cheng, *The CRISP theory of hippocampal function in episodic memory.*, eng, Frontiers in Neural Circuits **7** (2013) 88.

[56] S. Royer and D. Paré, *Conservation of total synaptic weight through balanced synaptic depression and potentiation*, Nature **422** (2003) 518.

[57] G. Turrigiano, *Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function.*, Cold Spring Harbor Perspectives in Biology **4** (1 2012) a005736.

[58] M. Letellier, F. Levet, O. Thoumine and Y. Goda, *Differential role of pre- and postsynaptic neurons in the activity-dependent control of synaptic strengths across dendrites*, PLoS Biology **17** (2019) e2006223, ed. by A. Bacci.

[59] H. G. Rey et al., *Single Neuron Coding of Identity in the Human Hippocampal Formation*, Current Biology **30** (2020) 1152.

[60] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment **2008** (2008) P10008.

[61] LaPlante, Roan and others, *bctpy v0.5.2: Brain Connectivity Toolbox for Python*, URL: %7Bhttps://github.com/aestrivex/bctpy%7D.

CHAPTER 4

Dynamical Learning of Dynamics

Experiments indicate that some neural networks are capable of learning new tasks without changing their synaptic connectivity. We propose a scheme for supervised learning of driven dynamical systems by networks with fixed weights. After appropriate pretraining with synaptic modifications, the networks adapt their dynamics to learn new tasks without changing synaptic connectivity. Our scheme may find applications in neuromorphic computing, where weight modification can be slow and costly.

This chapter is an adaptation with focus on my contributions of the article of the same title that was published in Physical Review Letters under the reference: Christian Klos, Yaroslav Felipe Kalle Kossio, Sven Goedeke, Aditya Gilra, and Raoul-Martin Memmesheimer, Phys. Rev. Lett. 125, 088103, https://doi.org/10.1103/PhysRevLett.125.088103.

4.1 Introduction

Acquisition of novel behavior through learning is generally assumed to rely on modification of synaptic weights and connectivity within the neural network [1, 2]. However, experiments indicate that learning may also happen without modification of synaptic wiring [3–7]. In particular, such synapse-independent learning was observed when animals had previously learned similar tasks [5–7]: In a group of mice potentiation of synapses expressing N-methyl-D-aspartate receptors was blocked either chemically or by a saturation protocol. This group performed significantly worse than the control in a water maze navigation task. However, if these mice were pretrained prior to blocking on a different water maze navigation task, their performance was comparable to the control group. Can properly pretrained networks learn novel tasks without synaptic weight modification?

A network of simple non-linear neurons with static weights is capable of approximating any bounded function or for a finite-time dynamical system. In principle, such static network can then approximate another network that is learning by changing its synaptic weights, together with its learning algorithm by its dynamics [8, 9]. In other words, a static network can perform dynamical learning. Training the static network to approximate a weight-learning network is a type of meta-learning or "learning to learn" [10–12]. Recent research on meta-learning [11, 12] focused primarily on learning of reinforcement

learning. Other research studied learning of supervised dynamical learning, usually with supervisory signal present even during testing [13–18].

Motivated by experiments suggesting that pretrained networks can learn without weight modification and by recent work on meta-learning, we propose a model network capable of such learning. Our networks can perform supervised learning of tasks without weight modification, if they were pretrained on samples from the same family of tasks before. This chapter will, in particular, focus on the learning of driven dynamical systems and subsequent teacher-free generation of their trajectories. We use a paradigm of reservoir computing, where only the output weights are trained [19, 20]: The reservoir, a recurrent neural network, is driven by inputs that perturb its internal state; the output is a simple linear transformation of the reservoir state. We first pretrain a network allowing weight modifications on representative samples from a family of dynamical systems. After pretraining, the weights are fixed, while the network can dynamically learn different unseen dynamical systems of the same family given supervisory input. After the dynamical learning, the network can generate the learned dynamics without supervision.

4.2 Network model

We use a recurrent network of 1000 rate-coupled neurons [21, 22] as our reservoir. The rate of neuron i, $r_i(t) = \tanh(x_i(t) + b_i)$, is a nonlinear function of its state $x_i(t)$ and offset b_i [20, 23, 24]. The offsets are uniformly distributed between -0.2 and 0.2. The reservoir network has two outputs: a signal $z(t)$, and a context $c(t)$, which are also continually fed back to the network to allow their autonomous generation [20], Fig. 4.1. After learning, $z(t)$ should generate the desired dynamics while $c(t)$ should index it. The network is provided with an external driving input $u(t)$, and during pretraining also with a supervisory error input $\varepsilon(t) = z(t) - \tilde{z}(t)$, indicating the difference between generated and target trajectories. For constant weights the network dynamics are given by

$$\begin{aligned} \tau\dot{x}(t) &= -x(t) + Ar(t) + w_z z(t) + w_c c(t) \\ &\quad + w_\varepsilon \varepsilon(t) + w_u u(t), \\ z(t) &= o_z r(t),\ c(t) = o_c r(t), \end{aligned} \tag{4.1}$$

where A is the recurrent weight matrix, τ is the (diagonal) matrix of time constants, o_z is signal and o_c is context output weights, w_z and w_c are the respective feedback weights, and w_ε, w_u are the error and drive input weights. The recurrent weights are set to zero with probability 0.8, and nonzero weights are drawn from a Gaussian distribution with mean 0 and variance $\frac{g^2}{0.2N}$, where $g = 1.5$ [23]. The input and the feedback weights are drawn from a uniform distribution between -2 and 2. The signal and context output weights are initially set to 0, these are the only plastic weights in the network.

4.3 Pretraining, dynamical learning and testing

The goal of pretraining is to allow network to learn dynamical systems from a certain family without weight modification while given only the error $\varepsilon(t)$ as a supervisory input. During the dynamical learning, the static network should associate unique constant context with the target dynamics. After fixing this context the learned dynamics should persist even after the supervisory input is removed.

The network is pretrained with weight modifications on representative target trajectories $\tilde{z}(t)$ of

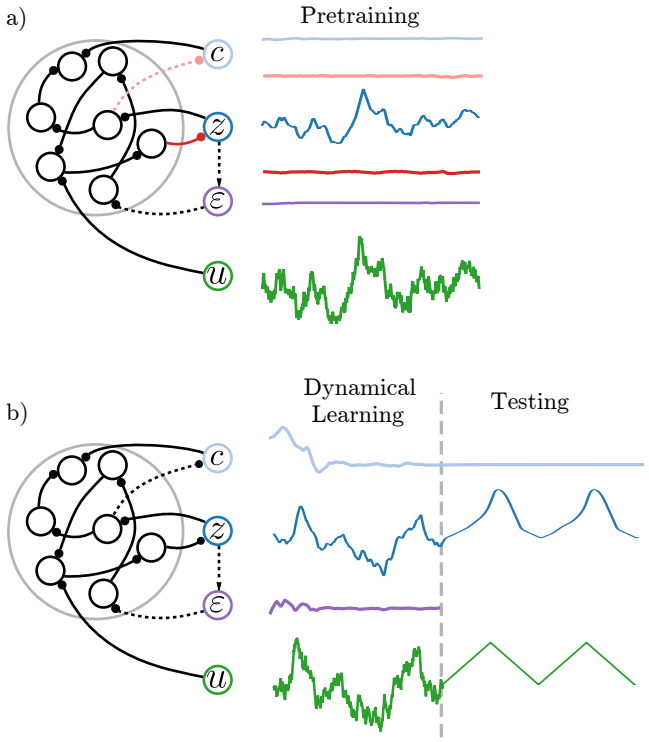

Figure 4.1: Schematics illustrating pretraining, dynamical learning and testing (traces are not to scale for clarity). (a) During pretraining, the network (gray circle) receives low-pass filtered white noise drive $u(t)$ (green) and corresponding supervisory signal $\varepsilon(t) = z(t) - \tilde{z}(t)$ (purple). The output weights (left, red and light red) of the network are modified with the FORCE rule using the errors $z(t) - \tilde{z}(t)$ and $c(t) - \tilde{c}(t)$ (right, red and light red), such that $z(t)$ and $c(t)$ (blue and light blue) match their targets. (b) Dynamical learning and testing. All the weights are now fixed. During dynamical learning, the network receives the supervisory signal $\varepsilon(t) = z(t) - \tilde{z}(t)$ and the corresponding low-pass filtered white noise $u(t)$ as inputs (purple and green), and adapts its dynamics to generate $z(t) \approx \tilde{z}(t)$ (blue). During testing, there is no supervisory signal and the context $c(t)$ is fixed to its previous average. The drive $u(t)$ is now a triangular wave.

driven dynamical systems from a particular family and the corresponding random realizations of drive $u(t)$. At the same time the network is provided with chosen target constant indexes $\tilde{c}$ that index the family. The target trajectories, corresponding drives, and contexts are presented for periods of $t_{\text{stay}} = 1000$, in a randomly repeating sequence for the total pretraining duration of $t_{\text{wlearn}} = 30000$. During pretraining the network always receives error feedback, $\varepsilon(t) = z(t) - \tilde{z}(t)$.

During the pretraining the output weights to $z(t)$ and $c(t)$ are trained with supervised recursive least-squares algorithm called FORCE [23] (with FORCE parameter α set to 1), Fig. 4.1a. The algorithm provides fast learning, which ensures that the network output and thus the feedback match the targets with only small errors. The network is then receiving largely correct feedback during training; the small deviations are generated intrinsically and resemble those that occur during testing. This allows the network to learn the correct dynamics in the presence of realistic fluctuations.

After pretraining is complete, all the weights are fixed, and the network can only learn new dynamical systems by its dynamics. The new dynamical systems were previously unseen, but come from the same family as the ones used for pretraining, Fig. 4.1b. The networks are presented with drive $u(t)$ and

the corresponding error $\varepsilon(t) = z(t) - \tilde{z}(t)$. The dynamical learning time $t_{\text{learn}} = 200$ is relatively short compared to the timescales of the dynamical system. For testing, the supervisory input is turned off, $\varepsilon(t) = 0$, and context $c(t)$ is fixed to a constant value, an average of ones assumed during dynamical learning, $c(t) = \bar{c}$. .

4.4 Results

We illustrate our approach by dynamically learning trajectories of driven overdamped pendulums with different masses, Fig. 4.2. These driven dynamical systems (pendulums) come from a family

$$\begin{aligned}\dot{\tilde{z}}(t) &= F(\tilde{z}(t), u(t); m)\\ &= -m\sin(\tilde{z}(t)) + u(t)\\ &\quad - \exp((\tilde{z}(t) - 0.65\pi)/0.65\pi) + \exp(-(\tilde{z}(t) + 0.65\pi)/0.65\pi),\end{aligned}$$

where $u(t)$ is a drive and m is the mass of a pendulum, the last two terms prevent full rotations of the pendulum. The network is pretrained on three pendulums with different masses, $m = 0.5, 1.0, 1.5$; the corresponding context targets are $\tilde{c} = 0.7, 0.95, 1.2$. Thereafter, a pendulum with an unseen mass ($m = 0.8$ or $m = 1.2$) is dynamically learned. Both pretraining and dynamical learning are based on imitation of trajectories. However, the network needs to generate unseen output trajectories during testing. As drive, during both pretraining and dynamical learning, we use low-pass filtered white noise $\dot{u}(t) = -u(t) + 0.2dW/dt$. This provides a variety of frequencies to comprehensively sample the dynamics of the pendulum. During testing, however, we use a qualitatively different drive: a triangular wave with unit amplitude and period $T = 50$. The network is able to approximate the completely novel target dynamics during testing, Fig. 4.2a,b. This shows that network must have learned the underlying vector field $F(\tilde{z}(t), u(t); m)$, and also that learning goes beyond interpolation of trajectories indicated by overlapping trajectories in Fig. 4.2b (blue and gray traces). To quantify the network performance, we measure the root-mean-square error between the target trajectory and the network output during testing, Fig. 4.2c. The error is small within and slightly beyond the range spanned by pendulums learned during pretraining.

4.5 Discussion and conclusion

We propose a scheme for teaching neural networks driven dynamical systems that does not require weight modifications during training. For this, networks are pretrained with weight modifications on the samples from the same family of dynamical systems. During part of the pretraining, the networks also receive supervisory input indicating the divergence from the desired dynamics. During the dynamical learning, error input allows networks to learn the correct dynamics and the corresponding context that indexes the family. During testing, when there is no supervisory signal, the fixed learned context input allows networks to exhibit the correct dynamics. Dynamical learning relies on this association between target dynamics and context.

Our work proposes an explanation to the observed learning without synaptic modifications [4–7]. After dynamical learning, the memory is stored in a constant activity of a single "context" neuron that index the task family. For short-term memory [25, 26] the task index can be kept in the activity.

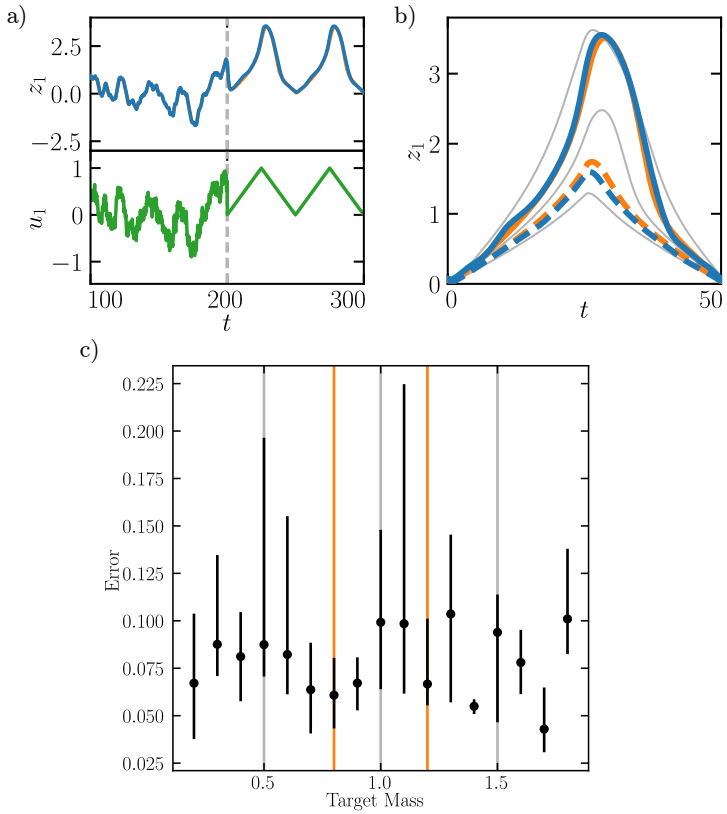

Figure 4.2: Dynamical learning and testing of a driven overdamped pendulum. (a) Drive (green), corresponding signal (blue) and target trajectory (orange) during dynamical learning (left of the dashed vertical) and subsequent testing (right of the dashed vertical). (b) Learned trajectories of two different pendulums (continuous and dashed blue) and coresponding target trajectories (continuous and dashed orange), driven by the same triangular wave input. Trajectories of pendulums learned during pretraining (grey). (c) Root-mean-square error between output and target trajectories. Data points show median error and error bars represent interquartile range between frist and third quartiles, calculated from ten network instances. Masses of the pretrained targets (grey vertical lines), masses of dynamically learned targets in (b) (orage vertical lines).

For longer-term storage, the memory may be consolidated by “context” neurons via internal cellular mechanisms, which may be similar to ones described in Ref. [27].

Our scheme may be applicable in neuromorphic computing where the plastic weights are typically difficult to realize and relatively slow to modify [28]. Our scheme may especially speed up learning in photonic reservoir computing, where weights are externally set to generate desired output dynamics [29–31].

References

[1] J. C. Magee and C. Grienberger, *Synaptic Plasticity Forms and Functions*, Annual Review of Neuroscience **43** (2020) 95.

[2] A. Citri and R. C. Malenka, *Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms*, Neuropsychopharmacology **33** (2008) 18.

[3] S. Chen et al., *Reinstatement of long-term memory following erasure of its behavioral and synaptic expression in Aplysia*, eLife **3** (2014) e03896, ed. by M. Ramaswami.

[4] M. G. Perich, J. A. Gallego and L. E. Miller, *A Neural Population Mechanism for Rapid Learning*, Neuron **100** (2018) 964.

[5] D. M. Bannerman, M. A. Good, S. P. Butcher, M. Ramsay and R. G. M. Morris, *Distinct components of spatial learning revealed by prior training and NMDA receptor blockade*, Nature **378** (1995) 182.

[6] D. Saucier and D. P. Cain, *Spatial learning without NMDA receptor-dependent long-term potentiation*, Nature **378** (1995) 186.

[7] M. K. Otnæss, V. H. Brun, M.-B. Moser and E. I. Moser, *Pretraining Prevents Spatial Learning Impairment after Saturation of Hippocampal Long-Term Potentiation*, Journal of Neuroscience **19** (1999) RC49.

[8] N. E. Cotter and P. R. Conwell, "Fixed-weight Networks Can Learn", *1990 IJCNN International Joint Conference on Neural Networks*, 1990 553.

[9] N. E. Cotter and P. R. Conwell, "Learning Algorithms and Fixed Dynamics", *IJCNN-91-Seattle International Joint Conference on Neural Networks*, 1991 799.

[10] S. Thrun and L. Pratt, eds., *Learning to Learn*, Springer US, 1998, ISBN: 978-0-7923-8047-4.

[11] J. Vanschoren, *Meta-Learning: A Survey*, arXiv:1810.03548 (2018).

[12] B. J. Lansdell and K. P. Kording, *Towards Learning-to-learn*, arXiv:1811.00231 (2018).

[13] L. A. Feldkamp, G. V. Puskorius and P. C. Moore, "Adaptation from Fixed Weight Dynamic Networks", *Proceedings of International Conference on Neural Networks (ICNN'96)*, vol. 1, 1996 155.

[14] L. A. Feldkamp, G. V. Puskorius and P. C. Moore, *Adaptive Behavior from Fixed Weight Networks*, Information Sciences **98** (1997) 217.

[15] A. S. Younger, P. R. Conwell and N. E. Cotter, *Fixed-weight On-line Learning*, IEEE Transactions on Neural Networks **10** (1999) 272.

[16] R. A. Santiago, "Context Discerning Multifunction Networks: Reformulating Fixed Weight Neural Networks", *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, vol. 1, 2004 189.

[17] M. Lukosevicius, *Echo State Networks with Trained Feedbacks*, tech. rep., Jacobs University Bremen, 2007.

[18] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein and W. Maass, *Long Short-term Memory and Learning-to-learn in Networks of Spiking Neurons*, arXiv:1803.09574 (2018).

[19] W. Maass, T. Natschläger and H. Markram, *Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations*,
Neural Computation **14** (2002) 2531.

[20] H. Jaeger and H. Haas, *Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication*, Science **304** (2004) 78.

[21] P. Dayan and L. Abbott,
Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems,
Cambridge: MIT Press, 2001.

[22] W. Gerstner, W. M. Kistler, R. Naud and L. Paninski,
Neuronal Dynamics - From single neurons to networks and models of cognition,
Cambridge: Cambridge University Press, 2014.

[23] D. Sussillo and L. F. Abbott,
Generating coherent patterns of activity from chaotic neural networks., eng,
Neuron **63** (2009) 544.

[24] M. Lukosevicius, H. Jaeger and B. Schrauwen, *Reservoir computing trends*,
KI - Künstliche Intelligenz **26** (2012) 365.

[25] K. K. Sreenivasan and M. D'Esposito, *The what, where and how of delay activity*,
Nature Reviews Neuroscience **20** (2019) 466.

[26] K. Oberauer, *Is Rehearsal an Effective Maintenance Strategy for Working Memory?*,
Trends in Cognitive Sciences **23** (2019) 798.

[27] F. Johansson, D.-A. Jirenhed, A. Rasmussen, R. Zucca and G. Hesslow,
Memory trace and timing mechanism localized to cerebellar Purkinje cells,
Proceedings of the National Academy of Sciences **111** (2014) 14930.

[28] E. Chicca, F. Stefanini, C. Bartolozzi and G. Indiveri,
Neuromorphic electronic circuits for building autonomous cognitive systems,
Proceedings of the IEEE **102** (2014) 1367.

[29] F. Duport, A. Smerieri, A. Akrout, M. Haelterman and S. Massar,
Fully Analogue Photonic Reservoir Computer, Scientific Reports **6** (2016).

[30] A. N. Tait et al., *Neuromorphic Photonic Networks Using Silicon Photonic Weight Banks*,
Scientific Reports **7** (2017).

[31] P. Antonik, M. Haelterman and S. Massar, *Brain-inspired Photonic Signal Processor for Generating Periodic Patterns and Emulating Chaotic Systems*,
Physical Review Applied **7** (2017).

CHAPTER 5

Summary

The synaptic wiring is assumed to be the main determinant of a network function. This thesis examines how simple plasticity rules can shape the wiring of the neural network. In particular, how wiring that leads to interesting dynamics may emerge and be preserved in the presence of noise. In Chapter 2 we demonstrate how homeostatic plasticity can drive neural networks close to the critical branching point where dynamics is characterized by neuronal avalanches. In Chapters 3 we show how Hebbian and homeostatic plasticity can maintain neuronal assemblies and the memories they represent despite a complete network rewiring and change of assembly neurons. Chapter 4 goes beyond synaptic plasticity, addressing types of learning that were suggested to not require synaptic modifications, and proposes a model for such learning.

In the second chapter, we show that simple homeostatic plasticity during development can allow a neural network to self-organize and approach a critical point. At the critical point, the long range interactions between neurons emerge from the short range interactions between directly connected neurons, leading to neuronal avalanches. We find that the requirement on the homeostatic plasticity is straightforward: it should drive the neural network to spike at a much higher rate than the spontaneous spiking rate of individual neurons. A further requirement is on the neural network: it should have a linear response to the spikes independent of its state. This ensures that the branching process does not depend on the state of the network, and all spikes always have the same average number of spikes as their offspring. The simplicity of the model allows us to show analytically that the network has the avalanche size and duration distributions with power law tails, as suggested by experiments. Our analysis also predicts the branching parameter (average offspring of a single spike) from easily measurable quantities: background spiking rate and induced spiking rate. Neuronal avalanches often overlap both in space (neurons) and time, making it hard to separate them in neuronal recordings. The simplicity of our model may stimulate the development of better methods for dealing with overlapping avalanches. Our model relies on the linearity of spike generation; biological neural networks are, however, not linear in all regimes. Future work should examine criticality and deviations from it when networks are in non-linear regimes. Finally, the critical branching processes we examine appear in diverse areas of research and our results, especially the analytical ones, may find applications there.

In the third chapter we examine how simple Hebbian plasticity rules together with homeostatic normalization can drive the drift of neuronal assemblies while at the same time preserving their general representational structure and memory. Our model suggests a solution to the conundrum of how memory can persist despite constant and substantial changes in network wiring and neural

representations. In our model, the wiring changes are gradual and representations are discrete, they therefore can be corrected with a simple "majority vote" implemented via Hebbian and homeostatic plasticity rules. There is no bound on the amount of drift: in principle all neurons in an assembly can be replaced and all synapses in the network change. The input and output neurons, however, always stay connected with an assembly and allow for a stable memory recall. Our scheme relies on slow changes and on discrete encoding of memories. It would be interesting to extend the model to continuous structures and representations, for example neural manifolds. Importantly, our simple mechanism for drift compensation may further the development of stable long-term brain-computer interfaces. Unfortunately, at present most of the available neural recordings either have too little neurons, too large intervals between recording sessions, or do not record the spontaneous network activity to allow testing of our error-correction mechanism.

The fourth chapter suggests that modifications of the synaptic wiring may not be necessary to learn new tasks if a similar set of tasks was previously learned by a network. In our model, appropriately pretrained neural networks learn new tasks by modifying their dynamics while the synaptic wiring is fixed. In particular, for recall of simple tasks, only the constant activity of a few "context neurons" needs to be remembered by a network. This memory may remain in the dynamics for short term storage, or may be stored in a neuron intrinsic way for a longer time. An interesting, experimentally verifiable prediction of our model is the emergence of "context neurons" indexing task families. Our learning scheme can also be applied to photonic reservoir computing, where weight modifications are usually much slower than the system dynamics. Dynamical learning may thus unlock the true speed of such systems.

APPENDIX A

Supplementary Material for Chapter 2

A.1 Manipulation of neural excitability

To address the impact of a typical experimental manipulation on the dynamics of our networks, we change the neural excitability or, equivalently, the coupling strength via g. We find that after decreasing it, in the short term activity is subcritical, Figs. A.1a, A.2a. The neurites react to the resulting overall loss of input by outgrowth, which leads to strengthening of connections and finally to the recovery of the near-critical state. An increase of neural excitability leads to very strong activity, which is quickly overcompensated by a decrease of connection strengths within the course of a single large avalanche, Fig. A.1c. The system becomes subcritical and regains the near-critical state more slowly thereafter. In biological neural networks overly large spiking activity is prevented by refractoriness. When decreasing neural excitability, our network models incorporating refractoriness behave similar to those without, Figs. A.1b, A.2b. When increasing excitability, in the short term they display an excess of medium size and large avalanches, Fig. A.2c. The overlap sizes and coupling strengths decrease until the near-critical state is regained, Fig. A.1d. This happens faster than the adaptation from subcriticality due to the still large excess of spiking: in Fig. A.2c the distribution in green is already similar to that in gray, in Fig. A.2a,b the distributions in blue are markedly different from those in gray.

In agreement with our findings, experimental studies have shown that a global decrease of excitatory synaptic strengths (decrease of network excitability) leads to subcritical activity while a global decrease of inhibitory synaptic strengths (increase of network excitability) leads to supercritical behavior with an excess of large avalanches [1, 2].

A.2 Robustness of avalanche characteristics against changes in the spontaneous and saturation spike rates

The stationary state avalanche size and duration distributions are largely independent of the choice of f_0 and f_{sat}, Fig. A.3. They are practically critical whenever f_0 is small against f_{sat}. For all finite values of f_{sat} and non-zero values of f_0 there is ultimately an exponential cutoff, see Eq. (5) and Eqs. (9), (10) after expanding around $a(t) = 0$. The large range parameter scans in Fig. A.3 are greatly facilitated by our analytical formulas: They allow us to efficiently determine the avalanche characteristics for the

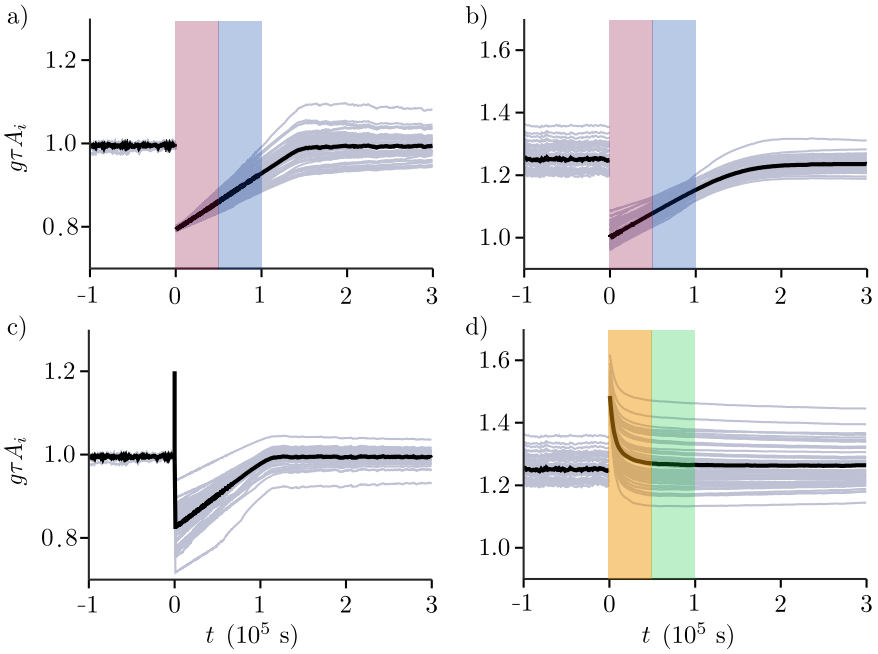

Figure A.1: Coupling strengths after manipulation of neural excitability. Scaled total overlaps of 50 neurons (gray) and mean total overlap (black), similar to Fig. 2c. At $t = 0$ s, the excitability g of all neurons is decreased, $g \to 0.8g$ (a,b), or increased, $g \to 1.2g$ (c,d). Before, the networks are in a stationary state. Neurons in (b,d) have refractory period $\tau_{\text{ref}} = \tau$. Avalanche size distributions for the color shaded areas are displayed in Fig. A.2 (sampling time 0.5×10^5 s each). The network growth rate is set to $K^{-1} = 2 \times 10^7$ s prior to manipulation to allow longer sampling times.

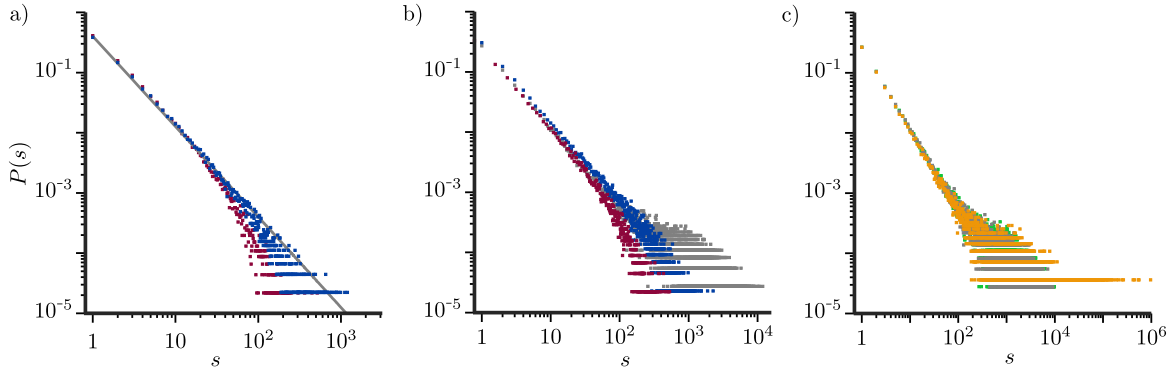

Figure A.2: Avalanche size distributions after manipulation of neural excitability. (a) Network of Fig. A.1a. Red and blue: distributions sampled directly after decreasing excitability [Fig. A.1a, red shading] and in the subsequent interval [Fig. A.1a, blue shading], respectively. Gray: analytical stationary state distribution Eq. (4). (b) Like (a) for the network of Fig. A.1b. Gray: stationary state distribution (sampled from simulation). (c) Network of Fig. A.1d. Orange: distribution sampled directly after increasing excitability [Fig. A.1d, orange shading]. Green: distribution sampled during the subsequent interval [Fig. A.1d, green shading]. Gray: stationary state distribution.

markedly subcritical as well as for the near-critical regime, where usually long numerical simulations are necessary to capture the distribution tails with their large and long avalanches. For illustration, Fig. A.3 exemplarily highlights the results for the parameters used in Fig. 3 as well as for parameter sets where f_{sat} is multiplied or divided by two (red, green, distributions are near-critical), or f_0 is increased by a factor of 50 (orange, distribution is markedly subcritical).

Values of f_0 and f_{sat} were directly measured in neural systems generating neuronal avalanches. The findings confirm that our assumption of $f_0/f_{\text{sat}} \ll 1$ is justified: Ref. [3] measures spike rates in retinal starburst amacrine cells in isolation and embedded in their network, where critical avalanches were reported a few days after birth [5]. The study finds spontaneous spike rates of isolated neurons that decrease from $f_0 \approx 1.3\,\text{mHz}$ to $f_0 \approx 0.1\,\text{mHz}$ between postnatal day 2 and postnatal day 6, Fig. 1 in Ref. [3]. The average spike rate $\bar{f} \approx 13\,\text{mHz}$ of connected neurons stays constant; we may assume $f_{\text{sat}} \approx 13\,\text{mHz}$ (or higher, if the homeostatic network plasticity is slow). The ratio f_0/f_{sat} therefore decreases from $f_0/f_{\text{sat}} \approx 0.1$ to $f_0/f_{\text{sat}} \approx 0.01$, Fig. A.3, cyan. Reference [4] investigates avalanches in neuronal cultures. The study reports a population spike rate, which rises from approximately 45 Hz at day 4 in vitro to approximately 730 Hz at day 30 in vitro, Fig. 5 in Ref. [4]. The first value is an upper estimate for $f_0 \times N$, where N is the number of neurons recorded from. The estimate neglects already existing couplings and mutual excitation between neurons after four days. The second value is a lower estimate for $f_{\text{sat}} \times N$, since the networks may grow further. We thus have $f_0 < 45\,\text{Hz}/N$ and $f_{\text{sat}} > 730\,\text{Hz}/N$, such that $f_0/f_{\text{sat}} < 0.06$, Fig. A.3, pink line.

A.3 Independence of avalanche characteristics of other model parameters

Equation (4) shows that the stationary state avalanche size distribution only depends on f_0 and f_{sat} (via $\sigma = 1 - f_0/f_{\text{sat}}$). Similarly, Eqs. (9), (10) show that the duration distribution depends only on f_0, f_{sat}, and τ. Equation (6) implies that changing τ is equivalent to rescaling the time axis and therefore only leads to a rescaling of the duration distribution. The dependence of avalanche statistics on the other model parameters, which do not enter Eqs. (4), (9), and (10), such as the coupling parameter g, the number of neurons N, the growth rate K, the dimensions of the square where the neurons are placed, and the way they are placed in (regular vs. random) vanishes during the network's self-organization process.

Consider as an example the coupling parameter g: If g is small, the network growth leads to larger total overlaps involving more synaptic partners. If g is large, the network growth leads to smaller total overlaps involving fewer synaptic partners. In the end, the dynamics are near-critical irrespective of g's value, Fig. A.4. The value of g may thus be chosen according to the biological system to which our model is applied. Indeed, ranges of overlaps and numbers of synaptic connection partners differ widely in full grown networks of potential relevance for our model: Neurons in cultures establish connections to tens and hundreds of other neurons, depending on the density of the plating [6], starburst amacrine cells receive inputs from tens of other starburst amacrine cells [5, 7], and in the intact cortex, neurons have thousands of different neurons as synaptic partners [8].

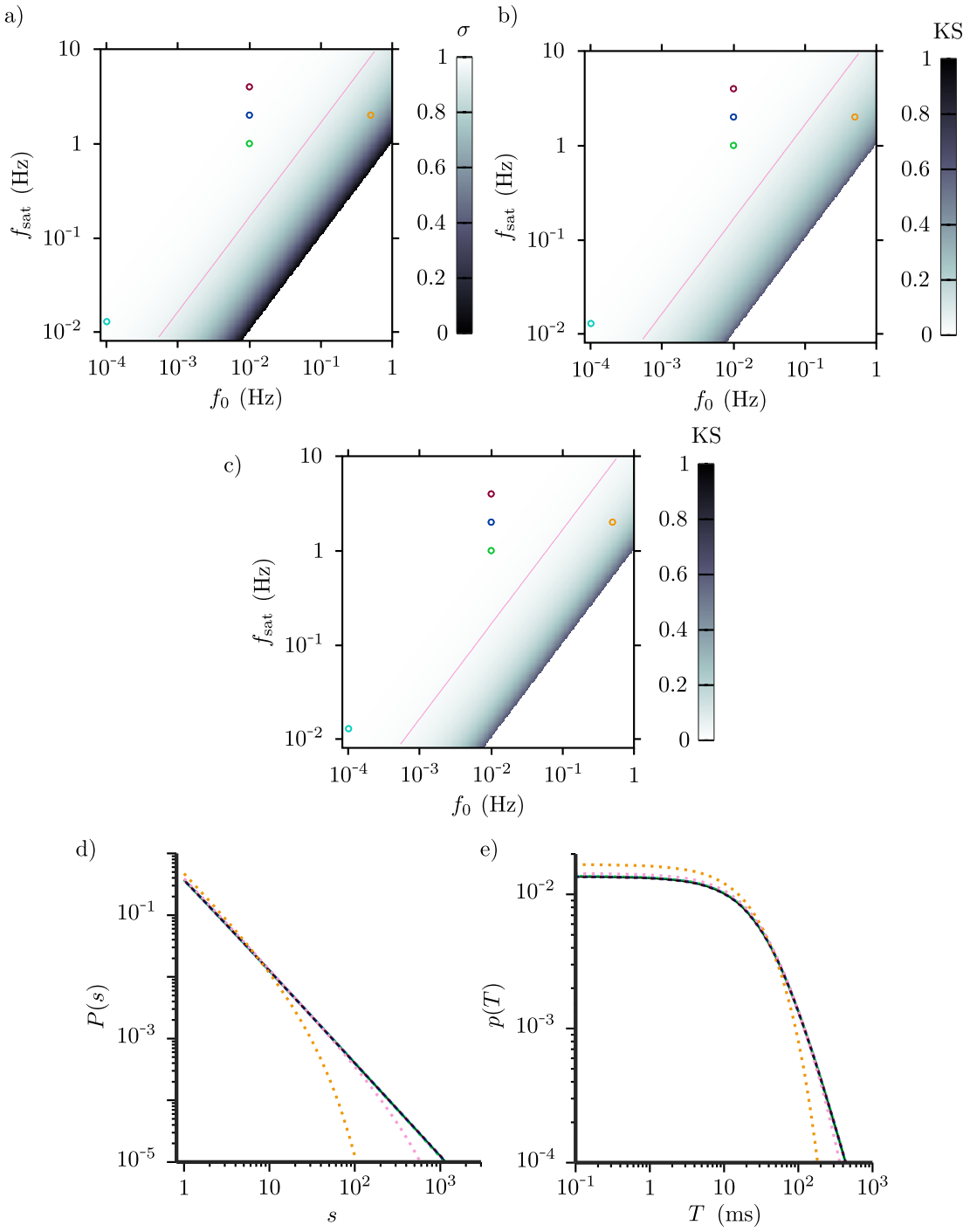

Figure A.3: Stationary state avalanche size and duration distributions are near-critical for large ranges of f_{sat} and f_0. (a) Branching parameter σ (gray scale, hatched area: $\sigma < 0$, system undefined) of the stationary state dynamics as a function of f_0 and f_{sat}. σ is close to 1 for large ranges of f_{sat} and f_0. (b) Kolmogorov-Smirnov distance (KS) between the critical size distribution [Eq. (4), $\sigma = 1$] and distributions with different f_{sat} and f_0 [Eq. (4), $\sigma = 1 - f_0/f_{sat}$]. The distance is close to 0 for large ranges of f_{sat} and f_0. Caption continues on next page.

Figure A.3: (c) Like (b) for the avalanche duration distributions Eqs. (9), (10). KS is close to 0 for large ranges of f_{sat} and f_0. Circles: $f_{\text{sat}} = 2\,\text{Hz}$, $f_0 = 0.01\,\text{Hz}$ (blue, $\sigma = 0.995$, used in the main text), $f_{\text{sat}} = 4\,\text{Hz}$, $f_0 = 0.01\,\text{Hz}$ (red, $\sigma = 0.9975$), $f_{\text{sat}} = 1\,\text{Hz}$, $f_0 = 0.01\,\text{Hz}$ (green, $\sigma = 0.99$), $f_{\text{sat}} = 13\,\text{mHz}$, $f_0 = 0.1\,\text{mHz}$ (cyan, $\sigma = 0.9923$) [3], $f_{\text{sat}} = 2\,\text{Hz}$, $f_0 = 0.5\,\text{Hz}$ (orange, $\sigma = 0.75$); pink line: $f_0/f_{\text{sat}} = 0.06$, upper limit of the ratio derived from Ref. [4] ($\sigma = 0.94$). (d,e) Size and duration distributions for the parameter values highlighted in (a-c) in alike colors (dotted for better discrimination, $\tau = 10\,\text{ms}$). The distributions in red, blue, green, cyan, and pink are near-critical, they partially overlay each other and the critical distributions (black). The distributions in orange are subcritical.

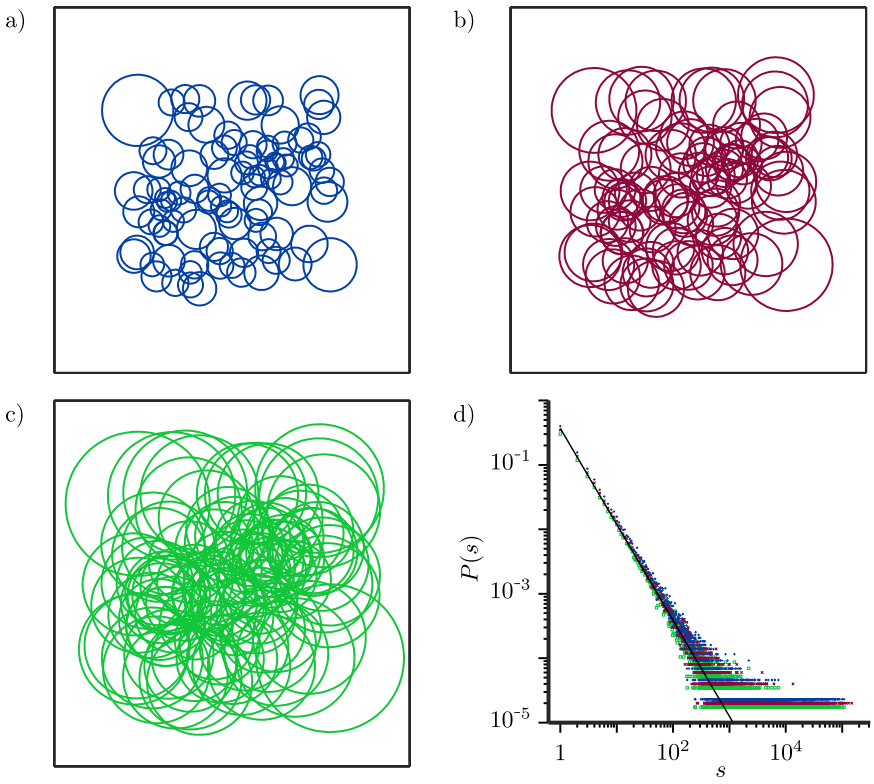

Figure A.4: The choice of the parameter g does not influence the avalanche statistics in the stationary state. (a), (b), and (c) display networks of $N = 100$ neurons in the stationary state for $g = 5\,\text{kHz}$, $g = 500\,\text{Hz}$, and $g = 50\,\text{Hz}$, respectively. (d) shows that the avalanche size distributions of these networks agree [networks in (a), (b), (c): blue +, red x, green squares]. Distributions are slightly vertically shifted for better discriminability.

A.4 Spontaneously active subpopulation

The stationary state avalanche size and duration distributions in our model are unchanged, if only a subpopulation of neurons is spontaneously active [9]. The quantity f_0 relevant for the statistics is the average spontaneous activity per neuron, see Fig. A.5. The size of the spontaneously active subpopulation may be small against N, which leads to an f_0 that is small against the individual spontaneous rates.

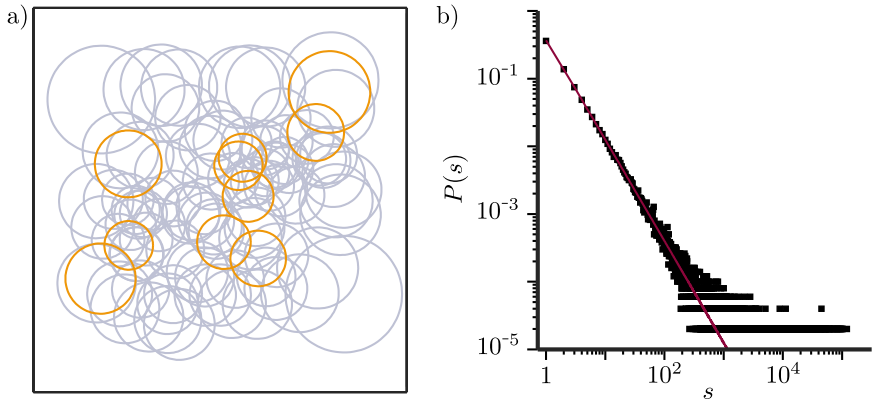

Figure A.5: Avalanche statistics in the stationary state are unchanged, if only a subpopulation of neurons is spontaneously active. (a) Network of neurons in the stationary state with 10 out of 100 neurons spontaneously active with rate $\tilde{f}_0 = 0.1$ Hz (orange). (b) The network's avalanche size distribution (numerically estimated, black) agrees with that of a network where all neurons are spontaneously active with $f_0 = 0.1\tilde{f}_0$ [Eq. (4), red].

A.5 Binning

We choose the bin size such that it keeps analytical probability estimates for different errors that are generated by binning moderate, Fig. A.6. The considered errors are: (i) joining the initial spike of an avalanche to the next avalanche, (ii) splitting the first spikes of the same avalanche, (iii) joining an average size avalanche to the next one, and (iv) splitting an average size avalanche. The resulting bin size t_{bin} depends on f_0, τ, N, and f_{sat}, which are experimentally accessible from single neuron measurements, anatomical data, and averaged spiking activity.

We first compute a simple estimate for the probability that binning joins the initial spontaneous progenitor spike of an avalanche to the next avalanche. Thereafter, we compute an estimate for the probability of splitting an avalanche between its first two spikes. These two probabilities yield a lower estimate for the probabilities of splitting and joining avalanches, as avalanches extending beyond their initial or first two spikes have higher probabilities of being joined to the next or being split. Keeping the obtained probabilities small provides an indication for an appropriate bin size, in particular because small avalanches are frequent. To compute the probability of joining the first spike of an avalanche to the next avalanche, we use that the rate of spontaneous progenitor spikes in the network is Nf_0. The interspike-interval (ISI) distribution of spontaneous spikes is therefore $p_{\text{ISI}}(t) = Nf_0 e^{-Nf_0 t}$. The probability of joining the initial spike of an avalanche to the following initial spike of an avalanche is approximately (the bin will usually not start at the first avalanche's start) the probability that the ISI between progenitor spikes is less than t_{bin},

$$P(\text{join first}) \approx \int_0^{t_{\text{bin}}} p_{\text{ISI}}(t)dt = 1 - e^{-Nf_0 t_{\text{bin}}}. \tag{A.1}$$

We now estimate the probability of splitting the first two spikes of an avalanche. This is approximately the probability that the second spike of the avalanche will occur more than t_{bin} apart from the first, $P(\text{split first}) \approx P(t_{\text{bin}} < t_2 < \infty)$, where t_2 is time of the second spike. The first spike increases the

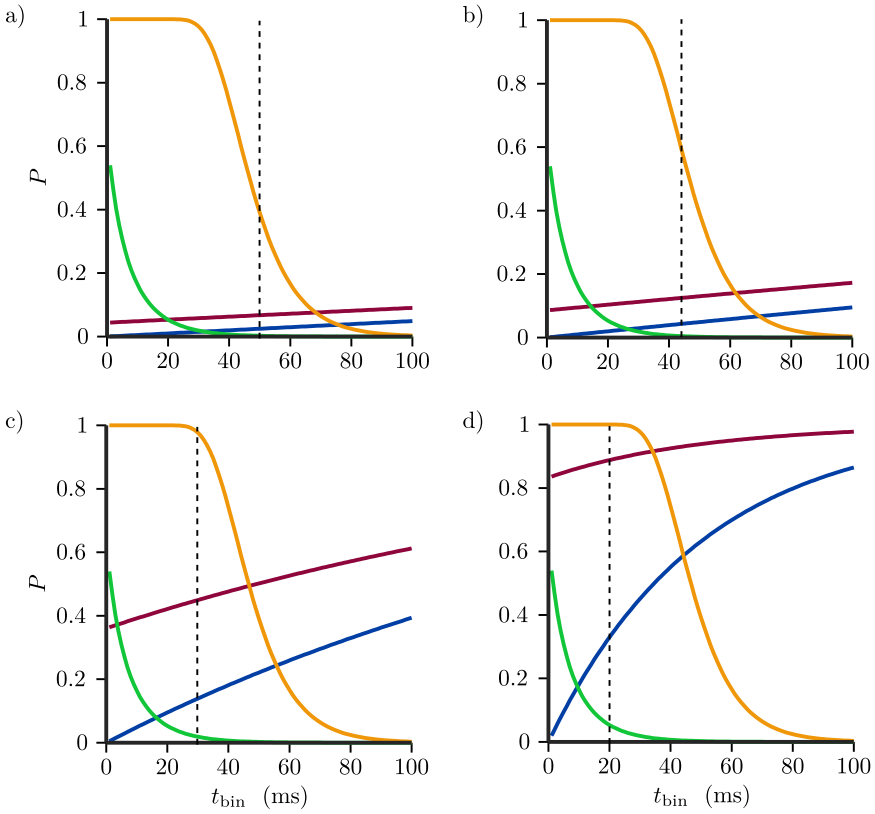

Figure A.6: Probability estimates for joining and splitting of avalanches for networks of (a) $N = 50$, (b) $N = 100$, (c) $N = 500$, (d) $N = 2000$ neurons and $\tau = 10\,\text{ms}$, $f_0 = 0.01\,\text{Hz}$, $f_{\text{sat}} = 2\,\text{Hz}$. Each panel displays our estimates for joining and splitting first spikes and average avalanches $P(\text{join first})$ Eq. (A.1) (blue), $P(\text{split first})$ Eq. (A.4) (green), $P(\text{join average})$ Eq. (A.5) (magenta), and $P(\text{split average})$ Eq. (A.6) (orange) versus bin size. The dashed line in (b) indicates the bin size chosen for the data analysis in Figs. 3, 4, A.2, A.4, and A.5; the other dashed lines indicate suitable bin sizes for other network sizes.

firing rate of the system by σ/τ, so $P(\text{split first})$ can be written as

$$P(\text{split first}) \approx P(t_{\text{bin}} < t_2 < \infty) = P(t_{\text{bin}} < t_2) - P(t_2 = \infty) \tag{A.2}$$

$$= e^{-\int_0^{t_{\text{bin}}} \frac{\sigma}{\tau} e^{-t/\tau} dt} - e^{-\int_0^{\infty} \frac{\sigma}{\tau} e^{-t/\tau} dt} \tag{A.3}$$

$$= e^{-\sigma(1-e^{t_{\text{bin}}/\tau})} - e^{-\sigma}. \tag{A.4}$$

To assess how to choose the bin size to keep probabilities of joining and splitting larger avalanches moderate as well, we derive similar estimates for avalanches with average length and duration. The probability of joining an avalanche of average duration $\bar{T}$ to the next one is approximately

$$P(\text{join average}) \approx \int_0^{\bar{T}+t_{\text{bin}}} p_{\text{ISI}}(t)dt = 1 - e^{-N f_0 (\bar{T}+t_{\text{bin}})}; \tag{A.5}$$

we compute $\bar{T}$ by numerically integrating $\bar{T} = \int_{0^-}^{\infty} T p(T) dT$, where $p(T)$ is given by Eq. (10). To

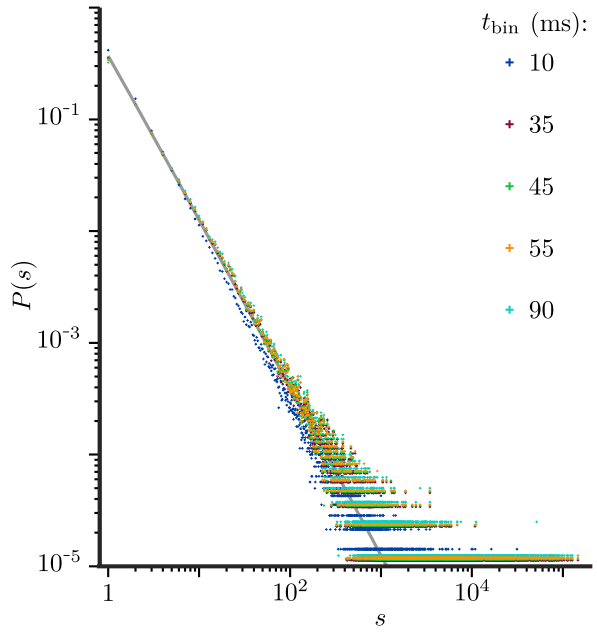

Figure A.7: Robustness against large changes in bin size. The figure displays avalanche size distributions obtained using bin sizes $t_{\mathrm{bin}} = 10$ ms (blue), $t_{\mathrm{bin}} = 35$ ms (red), $t_{\mathrm{bin}} = 45$ ms (green), $t_{\mathrm{bin}} = 55$ ms (orange), and $t_{\mathrm{bin}} = 90$ ms (cyan). The distributions obtained with $t_{\mathrm{bin}} = 35$ ms, $t_{\mathrm{bin}} = 45$ ms, $t_{\mathrm{bin}} = 55$ ms, and $t_{\mathrm{bin}} = 90$ ms are similar, the one obtained with $t_{\mathrm{bin}} = 10$ ms deviates, as expected from Fig. A.6b. The analyzed spike data are generated by a network with our standard parameters $N = 100$, $\tau = 10$ ms, $f_0 = 0.01$ Hz, $f_{\mathrm{sat}} = 2$ Hz; we usually use $t_{\mathrm{bin}} = 45$ ms for such data, Fig. A.6b.

estimate the probability of splitting an avalanche of approximately average size $\bar{s} = 1/(1-\sigma)$ [mean of the Borel distributed avalanche size, Eq. (4)], we first note that the split may occur about $\bar{s} - 1 = \sigma/(1-\sigma)$ times, which is the average number of offspring spikes in the branching process. We assume that the excitation from the previous avalanche spike has decayed to nearly zero when another one occurs, such that the next spike is generated by the previous only. Each spike of the avalanche then increases the collective firing rate to about σ/τ above the level of spontaneous spiking, like a progenitor spike. Since this implies that interspike-intervals are long, our assumption is conservative and gives us a higher probability to split the avalanche. It allows us to employ the already derived $P(\text{split first})$ as an estimate for splitting one of the $\sigma/(1-\sigma)$ intervals between avalanche spikes. The probability of not splitting an average size avalanche is approximately $(1 - P(\text{split first}))^{\sigma/(1-\sigma)}$ and thus

$$P(\text{split average}) \approx 1 - (1 - P(\text{split first}))^{\frac{\sigma}{1-\sigma}} \approx 1 - \left(1 - e^{-\sigma(1-e^{t_{\mathrm{bin}}/\tau})} - e^{-\sigma}\right)^{\frac{\sigma}{1-\sigma}} . \tag{A.6}$$

Replacing $P(\text{split first})$ by the probability of splitting an avalanche of two spikes yields similar results.

Figure A.6 displays the four probabilities Eqs. (A.1), (A.4), (A.5), and (A.6) against bin size. For small bin size there is a high probability of splitting an avalanche, which would result in overestimating the decay of avalanche distributions and possible critical exponents. The probability of splitting an avalanche becomes negligible for large bin size. For large bin size, however, there is a high probability

of joining avalanches, which would result in underestimating the decay of avalanche distributions and possible critical exponents. Our above estimates allow us to choose a bin size in between. For a faithful detection of the avalanche characteristics, it is more important to avoid joining small avalanches and splitting initial avalanche spikes, since they are most frequent. The bin size should thus be chosen such that keeping the probabilities $P(\text{join first})$ and $P(\text{split first})$ small is attributed a higher weight than keeping the probabilities $P(\text{join average})$ and $P(\text{split average})$ small. For the binning of our numerical data, we thus choose a bin size in the middle of the interval delimited by the crossings of $P(\text{join first})$ and $P(\text{split first})$ on the left and $P(\text{join average})$ and $P(\text{split average})$ on the right.

Our results are not sensitive to the chosen bin size. Since the time scales of avalanche dynamics and avalanche generation are well separated, most avalanches are relatively short and far apart, so the probability to join two avalanches increases slowly with bin size and there is a large range of suitable ones, see Fig. A.7.

References

[1] J. Beggs and D. Plenz, *Neuronal avalanches in neocortical circuits*, Journal of Neuroscience **23** (2003) 11167.

[2] W. L. Shew, H. Yang, S. Yu, R. Roy and D. Plenz, *Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches*, Journal of Neuroscience **31** (2011) 55.

[3] K. J. Ford, A. L. Félix and M. B. Feller, *Cellular Mechanisms Underlying Spatiotemporal Features of Cholinergic Retinal Waves*, Journal of Neuroscience **32** (2012) 850.

[4] Y. Yada et al., *Development of neural population activity toward self-organized criticality.*, Neuroscience **343** (2017) 55.

[5] M. H. Hennig, C. Adams, D. Willshaw and E. Sernagor, *Early-Stage Waves in the Retinal Network Emerge Close to a Critical State Transition between Local and Global Functional Connectivity*, Journal of Neuroscience **29** (2009) 1077.

[6] J. Barral and A. D Reyes, *Synaptic scaling rule preserves excitatory–inhibitory balance and salient neuronal network dynamics*, Nature Neuroscience **19** (2016) 1690.

[7] J. Zheng, S. Lee and Z. Zhou, *A transient network of intrinsically bursting starburst cells underlies the generation of retinal waves*, Nature Neuroscience **9** (2006) 363.

[8] M. Abeles, *Corticonics: Neural circuits of the cerebral cortex*, Cambridge Univ. Press, 1991.

[9] S. Johansson and P. Arhem, *Single-channel currents trigger action potentials in small cultured hippocampal neurons.*, Proceedings of the National Academy of Sciences **91** (1994) 1761.

APPENDIX B

Supplementary Material for Chapter 3

B.1 Supplementary figures

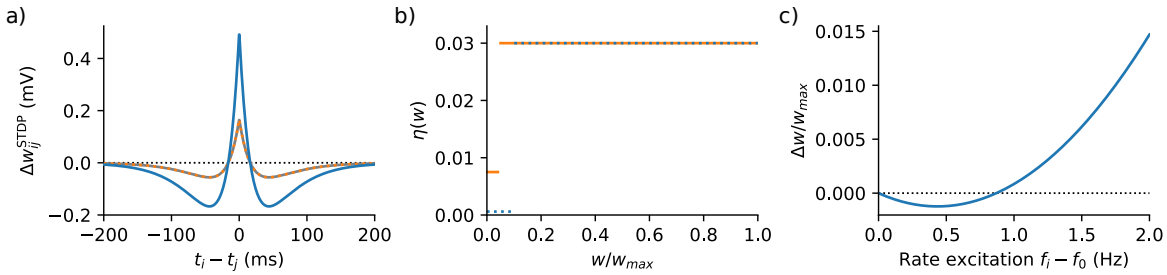

Figure B.1: Plasticity rules. (a) STDP windows used in the LIF model where noisy activity drives assembly drift (e.g. main text Fig. 2) for synapses between interior neurons (blue) and for synapses between interior and periphery neurons (dashed orange); STDP window used in the LIF model where spontaneous synaptic turnover drives assembly drift (Fig. 5) for all synapses (gray). The weight change is given in terms of the change of the peak EPSP that a synapse evokes in a resting neuron. (b) Weight dependence of the learning rate in the binary model (Fig. 4 middle, Fig. B.6), for interior (dashed blue) and for periphery (solid orange) neurons. (c) Dependence of the weight update on the excitation level in the Poisson model (Fig. 8c, Fig. B.13). Black dotted lines in (a) and (c) indicate border between depression and potentiation.

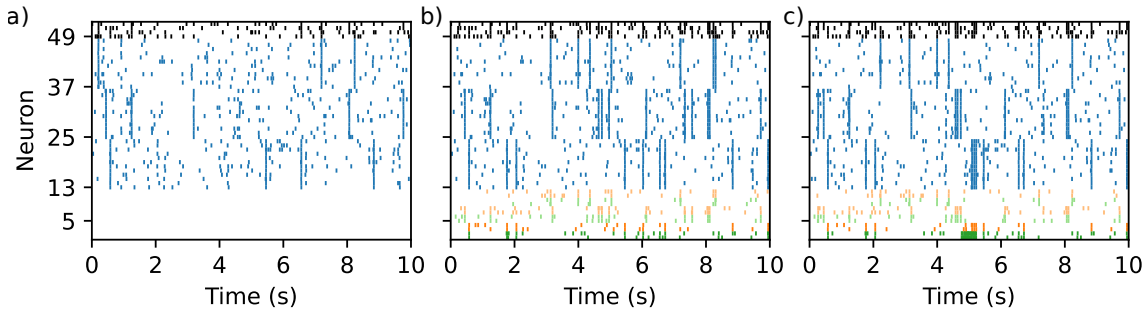

Figure B.2: Associative memory property and basic input-output functionality of assemblies. (a) Assemblies occasionally reactivate already due to the sparse background activity of the interior neuron population. (b) The rate of reactivation is higher if the activity of the periphery neurons is present. (c) Stimulation of a pair of input neurons (green, neurons 1,2) activates their assembly (assembly 1), which specifically activates its output neurons (orange, neurons 3,4), demonstrating basic functionality of our circuit. The spike trains are sorted according to the assemblies that the neurons belong to at $t = 0$s. The input neurons to assemblies 1,2 and 3 have indices 1,2 (green), 5,6 (light green) and 9,10 (light green). The output neurons of assemblies 1,2 and 3 have indices 3,4 (orange), 7,8 (light orange) and 11,12 (light orange). The first twelve assembly neurons of each assembly are displayed in blue, with indices 13-24, 25-36 and 37-48. Further, the spike trains of four inhibitory neurons are shown in black.

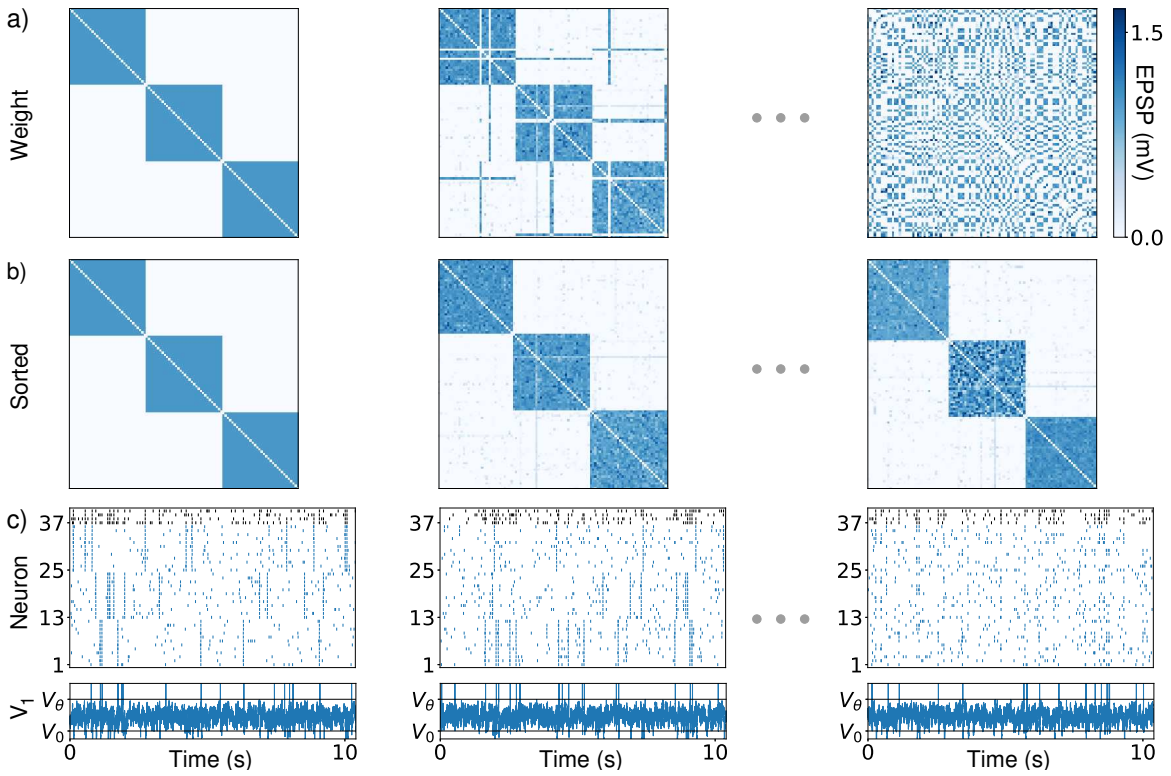

Figure B.3: Assembly drift in an LIF neuron network without periphery neurons. (a) Weights between neurons. First column: Network initialization with three assemblies. Second column, after 15 minutes: first transitions of neurons to a new assembly. Third column, after 12h: the assemblies have drifted away; the weight matrix has completely remodeled. (b) Like (a) but with neurons reordered according to assemblies that they belong to. The assembly structure is conserved over time. (c) Spike trains of 12 neurons from each of the ensembles that initially form assembly 1 (1-12), 2 (13-24) and 3 (25-36) and of four inhibitory neurons (black). (c, lower) membrane potential of neuron 1.

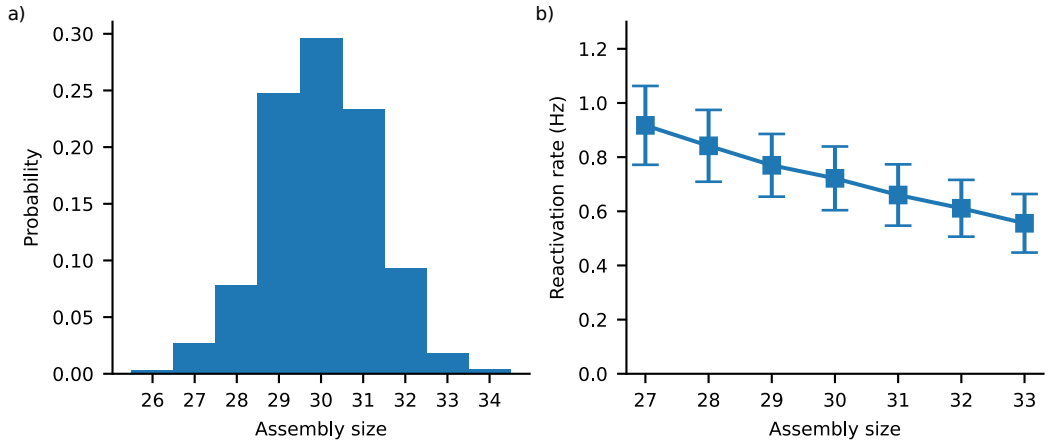

Figure B.4: Assembly size distribution and reactivation rate, for a LIF network as in main text Fig. 2. (a) Assembly sizes are unimodally distributed around their mean, $N_{\text{int}}/3 = 30$. Sizes are sampled every 270s during the 75h long simulation.(b) Assembly reactivation rate decreases with assembly size. We consider an event where the number of spikes emitted by assembly neurons within 15ms exceeds 50% of the assembly size as a reactivation (cf. Fig. 8). Reactivation rates are measured during 60s long intervals starting every 270s; squares show the mean, error bars the standard deviation over the measuring intervals. The panel displays reactivation rates for assembly sizes that are observed at least 50 times.
We explain the higher reactivation rate by the fact that the average synaptic weight of an assembly neuron to another one is larger in a small assembly such that less of a small assemblies' neurons have to be coincidentally active to initiate its reactivation. In larger assemblies there are more neurons that can be spontaneously active. However, because the smaller average input weight requires a larger number of neurons to be coincidentally active to initiate the reactivation, the exponential decay of the coincidence probability outweighs the larger number of possibilities. The larger average individual synaptic strength results from the fact that in a smaller assembly a neuron distributes its total output weight w_{sum} between less neurons, since it targets mainly neurons of the same assembly.

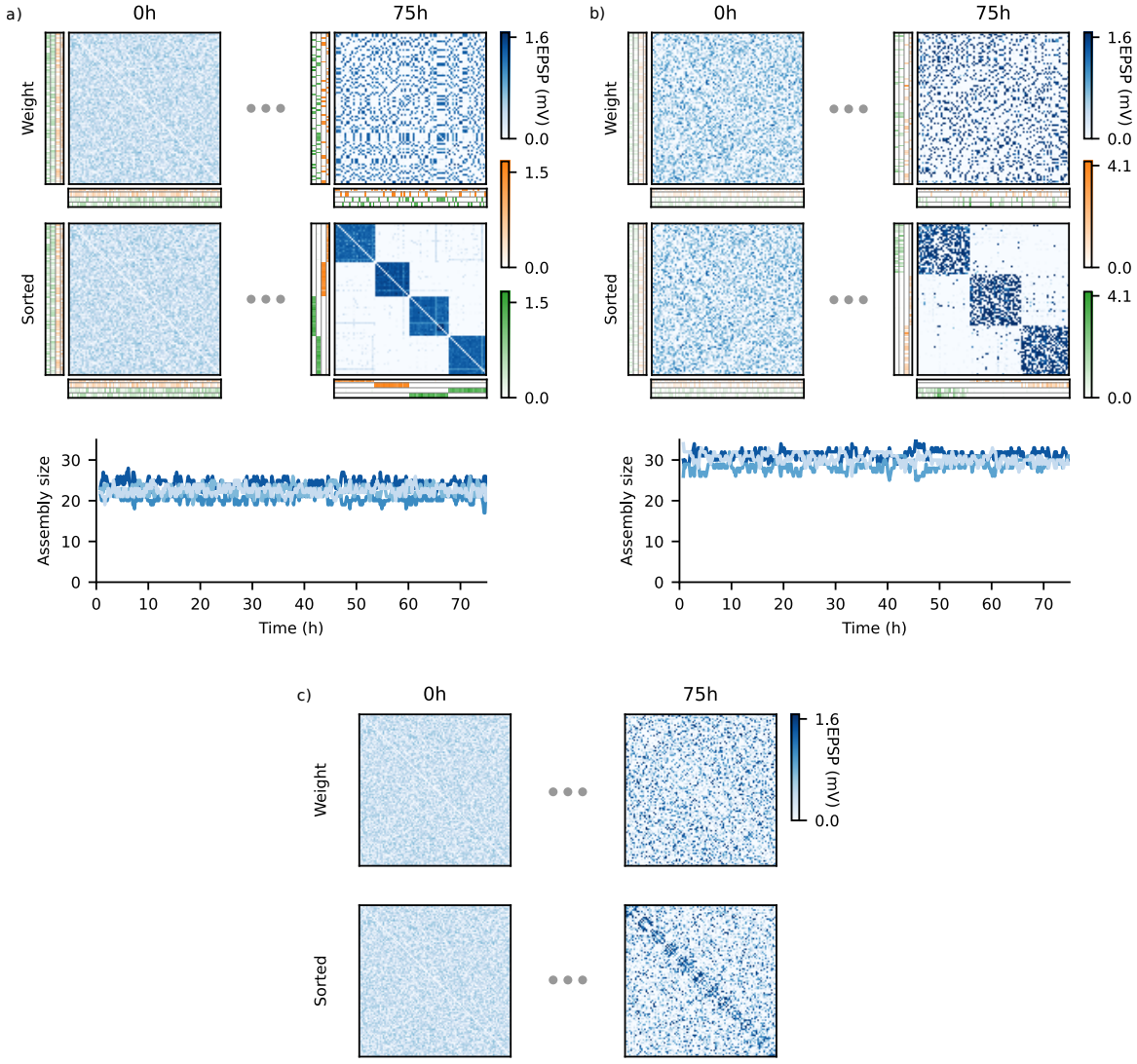

Figure B.5: See next page for caption.

Figure B.5: Spontaneous development of drifting assemblies. The figure displays results from networks of LIF neurons (a) where noisy autonomous activity drives assembly drift , (b) where spontaneous synaptic turnover drives assembly drift and (c) where noisy autonomous activity drives assembly drift and the network has no periphery neurons. Top parts of panels (a,b) and panel (c) are like in main text Fig. 2b,c. Bottom parts of panels (a,b) show the sizes of the emerged assemblies as a function of time. In contrast to our other network simulations, the networks are initialized by randomly drawing their excitatory weights from a uniform distribution and subsequently normalizing them (first columns). All other parameter values are unchanged. Within 70 minutes of simulated time, four assemblies have emerged in the networks shown in (a) and three drifting assemblies have emerged in the networks shown in (b). These assemblies persist and drift; the second columns show the network weights after 75h. The assembly sizes fluctuate (lower panels); mean assembly sizes differ from each other, due to the different numbers of periphery neurons. The fluctuations in the four assembly state (a) are larger than in the three assembly state in alike networks, cf. Fig. B.4. No clear assemblies emerge in the networks without periphery neurons where noisy autonomous activity drives assembly drift, panel (c). We made the same observations for five different realizations of each of the models. Panels (a) and (c) imply together with Figs. 2,5 that besides the dependence on parameters like the quotient $w_{\mathrm{sum}}/w_{\mathrm{max}}$ [1, 2], the number of assemblies in our models depends on the initial condition of the network. We expect that networks with the same general setup but quantitatively different parameters can also generate qualitatively different behavior with respect to the spontaneous emergence of assemblies. Finally, we note that the periphery neurons (if present) connect only randomly to the assemblies due to the random initial weights. Such networks may become functional for memory storage due to subsequent learning (see main text Discussion). We observed for one realization of the network with drift driven by noisy autonomous activity a switch of a periphery neuron from an assembly originally connected to four periphery neurons to an assembly originally connected to only two periphery neurons (at about 13.8h).

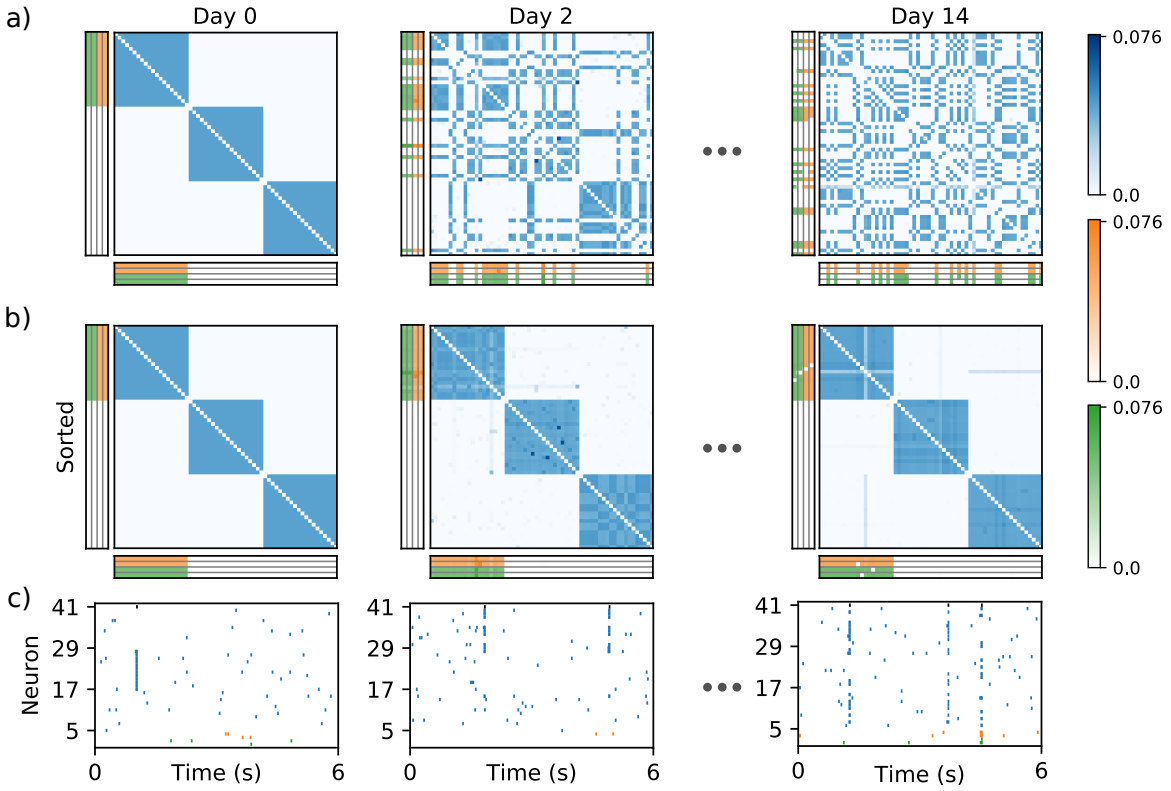

Figure B.6: Drifting assemblies in a network of binary neurons; noisy autonomous activity drives the drift. Display is like in main text Fig. 2. (a),(b) First column: Network initialization with three assemblies. Second column, after two days: several interior neurons switched to a new assembly. Third column, after 14 days: The assemblies have drifted away, the weight matrix has completely remodeled. (c) shows spike trains of the input (green) and output (orange) neurons of assembly 1, of 12 neurons from each of the ensembles that initially form assembly 1 (5-16), 2 (17-28) and 3 (29-40) and of the inhibitory unit (black).

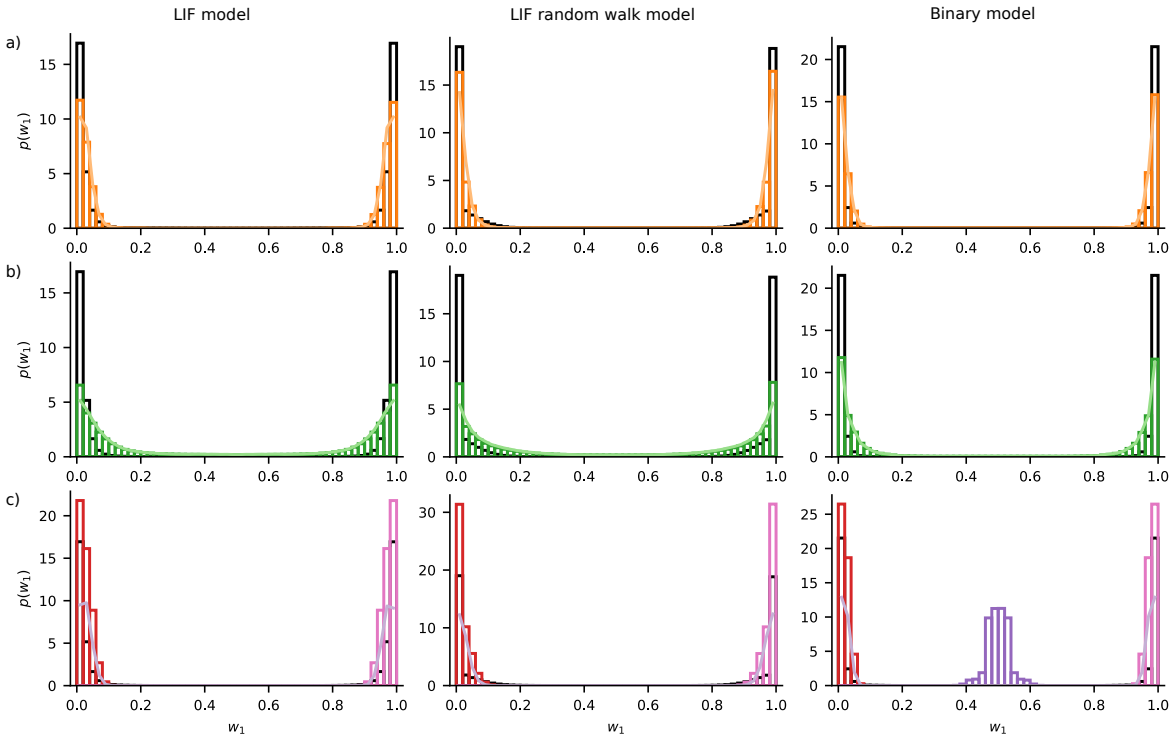

Figure B.7: Stationary distributions of w_1 for an LIF network (left, black histograms, network with two assemblies as in main text Fig. 4), the corresponding random walk model (middle, black histograms), and a binary network (right, black histograms, network with two assemblies as in Fig. 4), compared to the occupancy distributions of related Markov simulations (colored histograms) and an analytical diffusion approximation (colored curves). (a) Markov simulation accounting for drift and noise (orange) (b) Markov simulation accounting for noise only (green) (c) Markov simulation with drift and homogenized noise. For homogenized noise, the distributions of the Markov simulations depend on the initial w_1, since switching does not occur within the used simulation time and the neuron stays within one of the two (LIF, random walk) or three (binary) potential valleys. For the Markov simulation corresponding to the LIF and the corresponding random walk model the initial values are $w_1(0) = 0$ (red) and $w_1(0) = 1$ (pink). For the binary model they are $w_1(0) = 0$ (red), $w_1(0) = 1$ (pink), and $w_1(0) = 0.5$ (purple). Each of the resulting numerically estimated distribution parts with high occupancy probability is normalized to have unit integral. The analytical probability density distributions account for switching despite its low probability. Therefore they yield the different overall weighting of the high occupancy probability regions.

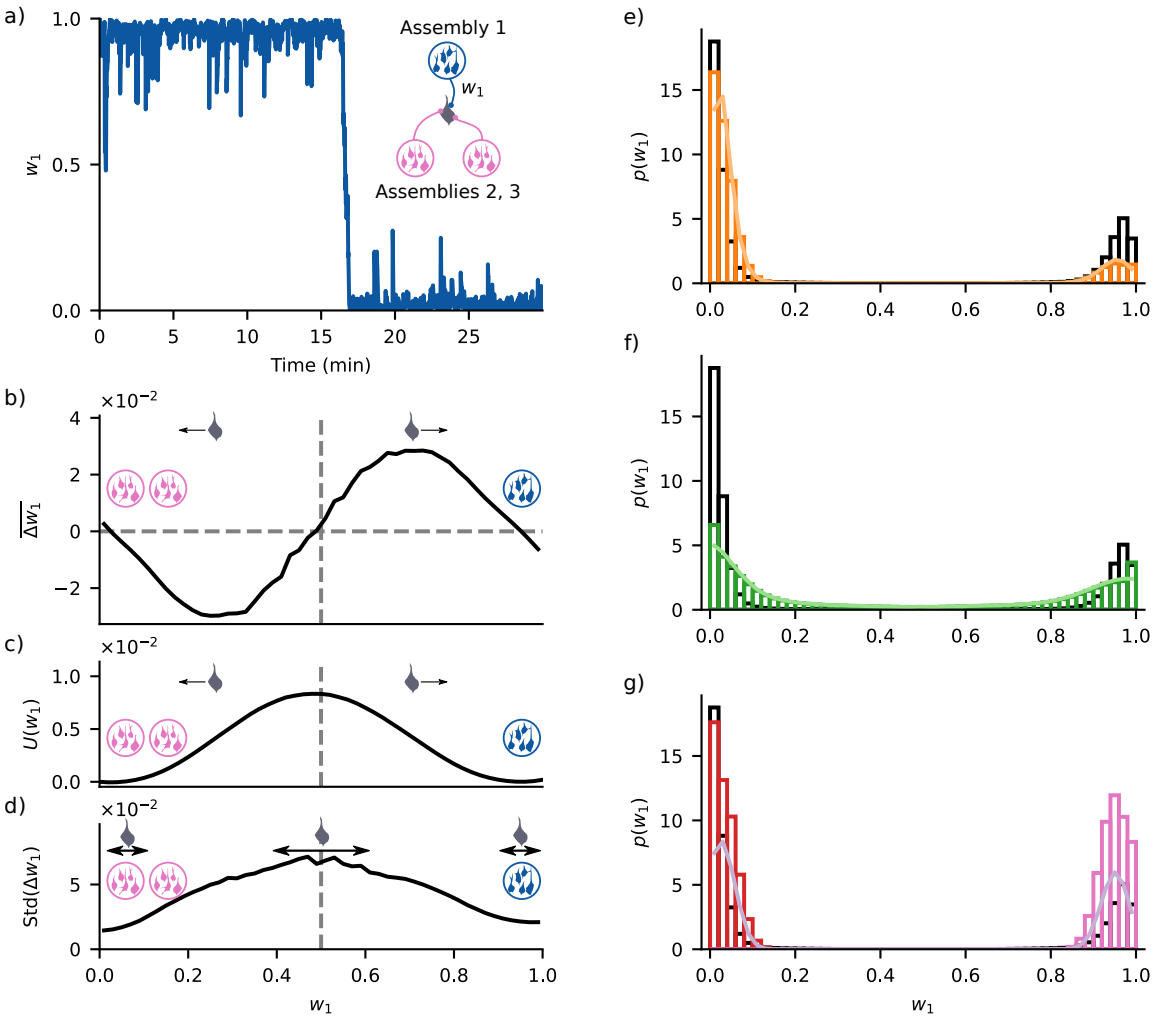

Figure B.8: Analysis of neuron transitions between assemblies and stationary distributions of w_1 for a LIF network with three assemblies. The network dynamics in (a-d) are those of Fig. B.3; display is like Fig. 4. Panels (e-g) are like Fig. B.7a-c. $w_1 \approx 1$ means that the test neuron belongs to assembly 1. $w_1 \approx 0$ means that it belongs to assembly 2 or 3 (or, rarely, that it is currently switching between them). Therefore the average weight changes (b), the potential (c), the noise strength (d) and the weighting of the high occupancy regions (e-f), which reflect the relative dwelling times, are markedly asymmetric (dashed lines in (b-d) inserted at $w_1 = 0.5$ and $\overline{\Delta w_1} = 0$ to guide the eye).

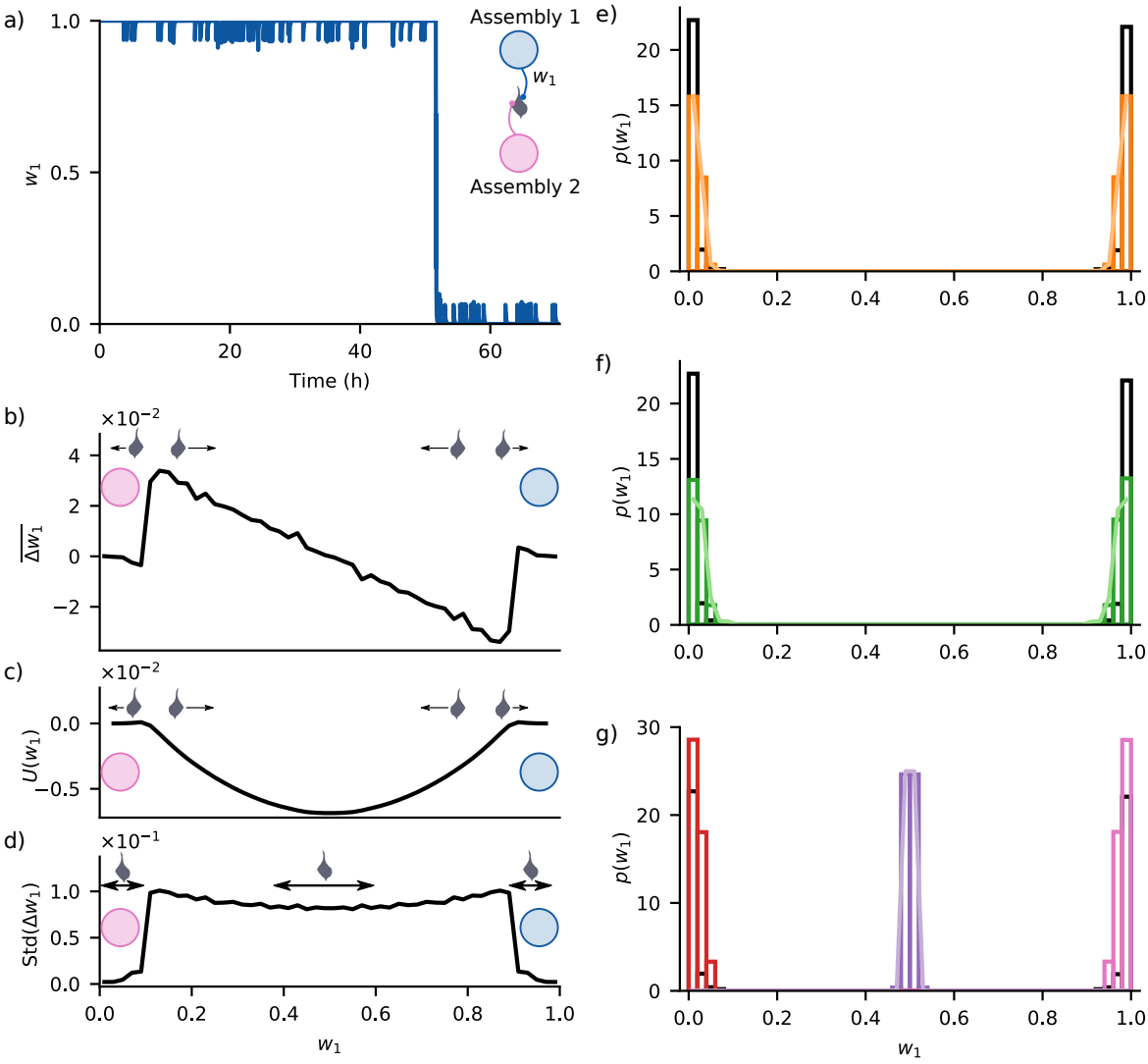

Figure B.9: Analysis of neuron transitions between assemblies and stationary distributions of w_1 for the binary random walk model. Display (a-d) is like main text Fig. 4, (e-g) like Fig. B.7a-c.

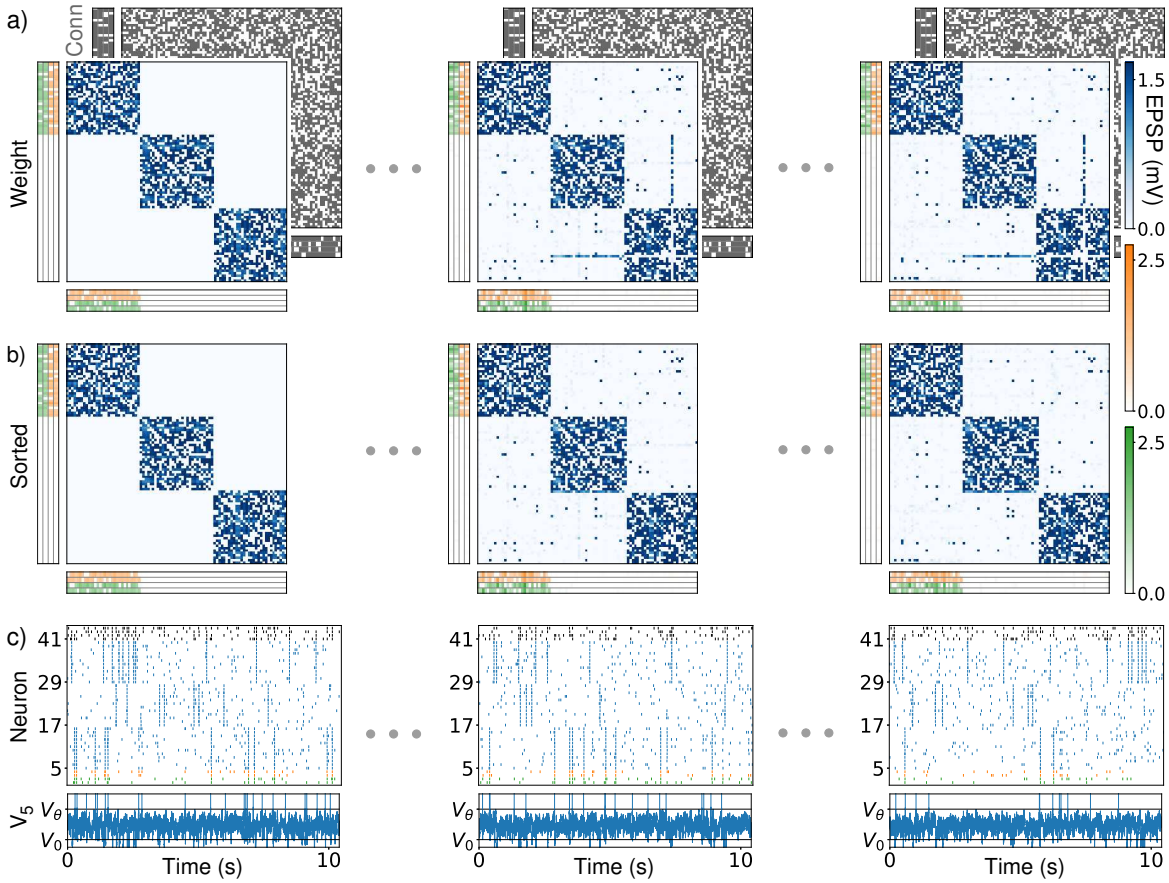

Figure B.10: LIF network as in main text Fig. 5 but without spontaneous synaptic turnover. Panels are like in Fig. 5, but for 0, 2 and 99 hours of simulated time. The connectivity matrix is constant. After an initial phase of adaptation (compare the first and second column), no more neurons change assembly (compare the second and third column).

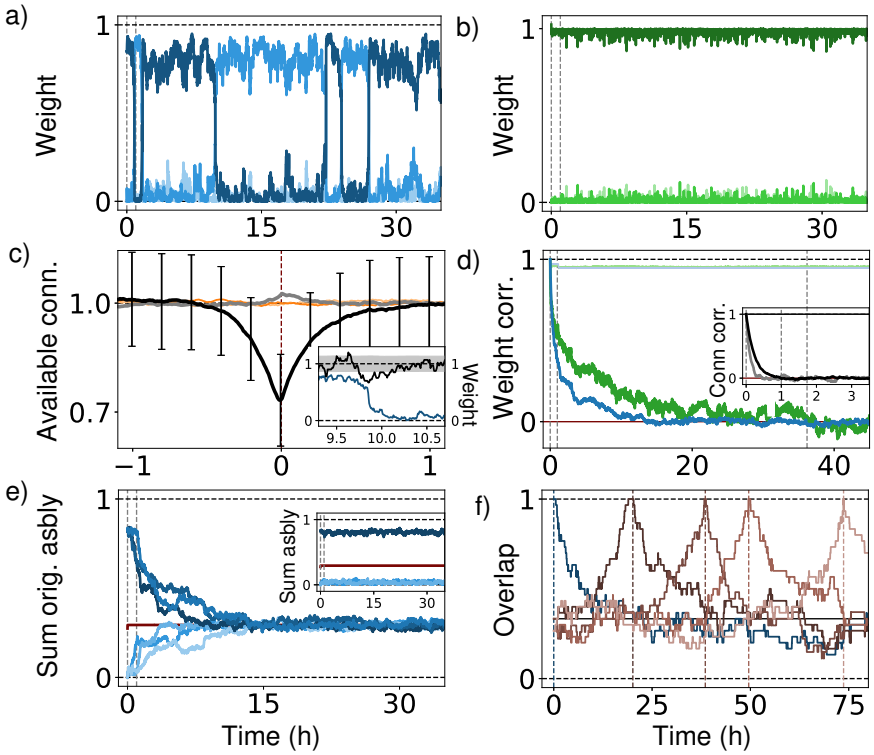

Figure B.11: Analysis of drifting assemblies and their periphery neurons for LIF networks where spontaneous synaptic turnover drives the drift. (a) Switching of the first interior neuron of main text Fig. 5 between the three assemblies, depiction like in Fig. 3b. Dashed black horizontals: maximal and minimal sums of weights. Gray dashed verticals: times used in Fig. 5. (b) Like (a) for the first periphery neuron. (c) Mechanism of switching. The displayed peri-switch time histogram shows a strong reduction of the number of input connections from the abandoned assembly near switching (black, error bars: standard deviations of distributions). The input connectivity from the entered assembly is slightly increased (gray). The output connectivity (orange, light orange) has no pronounced trend. The inset shows a typical switching event: available input connections (black; black dashed: average; gray: standard deviation) decay strongly before switching (weight from abandoned assembly: blue). To obtain the black curve in the peri-switch time histogram we collect the numbers of available input connections from the abandoned assembly and its periphery neurons to the switching neuron, for all switching events in the simulation of Fig. 5. We then normalize each number of available input connections by the expected number, using the current number of assembly neurons and the connection probabilities p_{int} and p_{peri}. Thereafter the collected pieces are centered at switching time, which is set to zero (red dashed vertical). The numbers of available input connections at a time point are then averaged and their standard deviation is computed. The other curves are computed likewise. The inset shows the third switching of assembly neuron 1, away from assembly 1, at about ten hours simulated time (cf. panel a). (d) Complete remodeling of network weights (main panel, like Fig. 3d) and connectivity (inset). Pearson correlations between initial and later weights of interior (blue) and periphery-interior synapses (green) converge to chance level. The same holds for the Pearson correlations between initial and later connectivity matrix entries of the interior (black) and periphery-interior (gray) connections The different decay times in main panel and inset show that synaptic weight plasticity compensates large parts of the spontaneous turnover. Networks with the same parameters but without synaptic turnover (main panel: light blue and green, inset: thin black and gray; partially overlapping) show an initial phase of adaptation to the underlying connectivity, where neurons that receive few input synapses from their assembly leave it (small jump-like decrease in light traces around one hour, cf. also Fig. B.10). Dashed horizontal: maximal correlation. (e) Complete assembly remodeling and maintenance of representational structure, like Fig. 3e. Dashed horizontals: maximal and minimal sums of weights. (f) Continued assembly drifting, like Fig. 3f. Dashed horizontals: bounds 0 and 1 of the overlap.

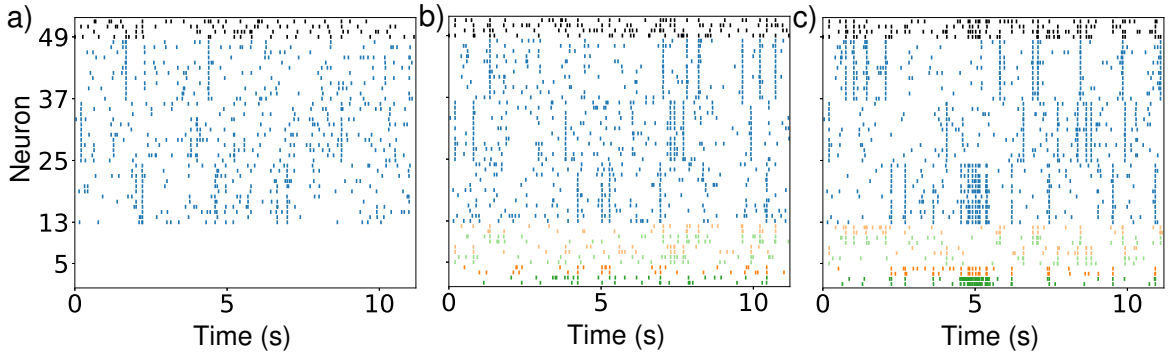

Figure B.12: Associative memory property and basic input-output functionality of assemblies for the network with drift driven by spontaneous synaptic turnover of main text Fig. 5. Description as for Fig. B.2.

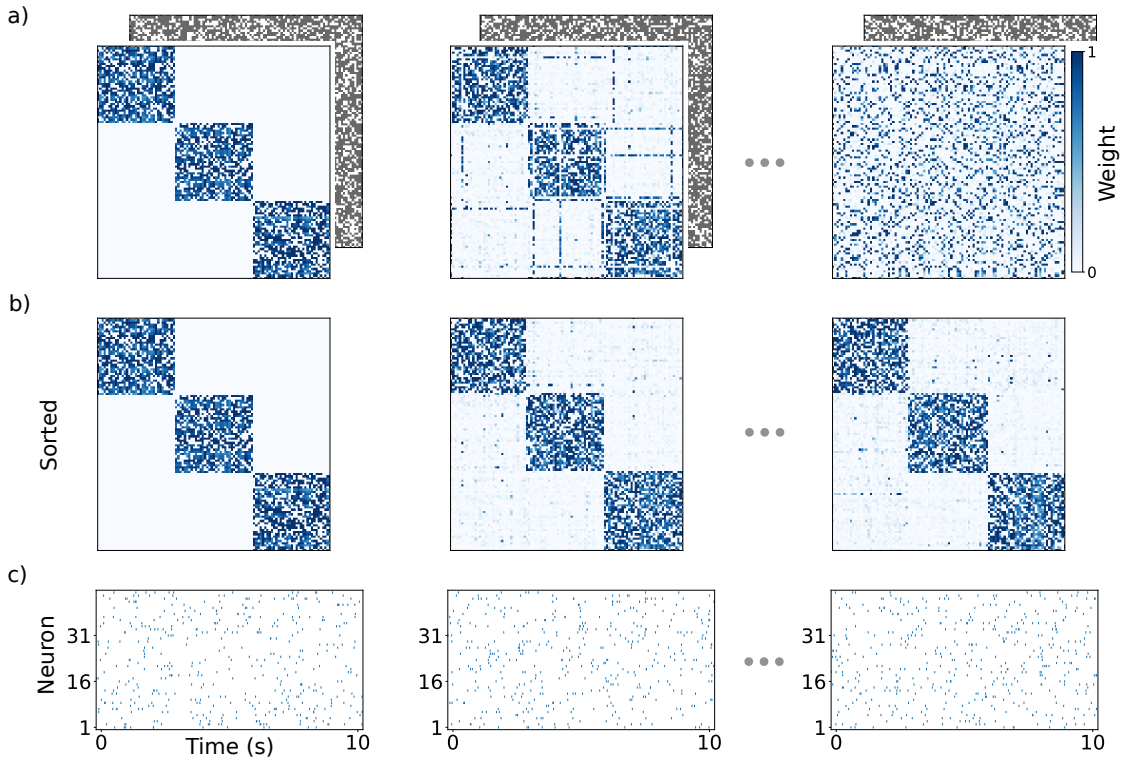

Figure B.13: Drifting assemblies in a network of linear Poisson spiking neurons with spontaneous synaptic turnover and without periphery neurons. Panels (a-c): like panels (b-d) in Fig. 2, but for 0, 72 and 1440 hours of simulated time, i.e. the second column shows the network after three days and the third column after 60 days. The underlying synaptic connections spontaneously turn over, which drives the drift. The weights are normalized by their maximal possible value $w_{\max}$. (c) Spike trains of 15 neurons from each of the ensembles that initially form assembly 1 (1-15), 2 (16-30) and 3 (31-45). Spiking activity is asynchronous and irregular without visible assembly reactivation.

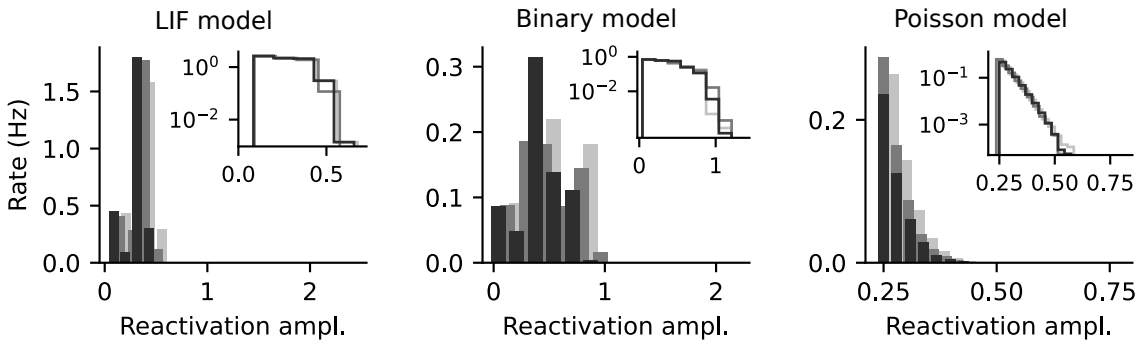

Figure B.14: Distributions of reactivation amplitudes in our network models as in main text Fig. 8c, but for randomly selected neuron ensembles of the same size as the three assemblies in the networks. For the LIF (left) and the binary (middle) model the histograms consist of amplitudes of about $1/3$ of the typical assembly reactivation amplitudes visible in Fig. 8c. This fraction agrees with the expected neuron overlap of the random ensembles with the assemblies. The observation therefore confirms the confinement of reactivations to single assemblies. For the linear Poisson network model (right) the amplitude distribution appears to exponentially decay in a similar fashion as for the assemblies in the network (Fig. 8c right). The semi-logarithmic plot of the complementary cumulative distribution in the inset reveals that for randomly selected ensembles the exponential decay is slightly faster.

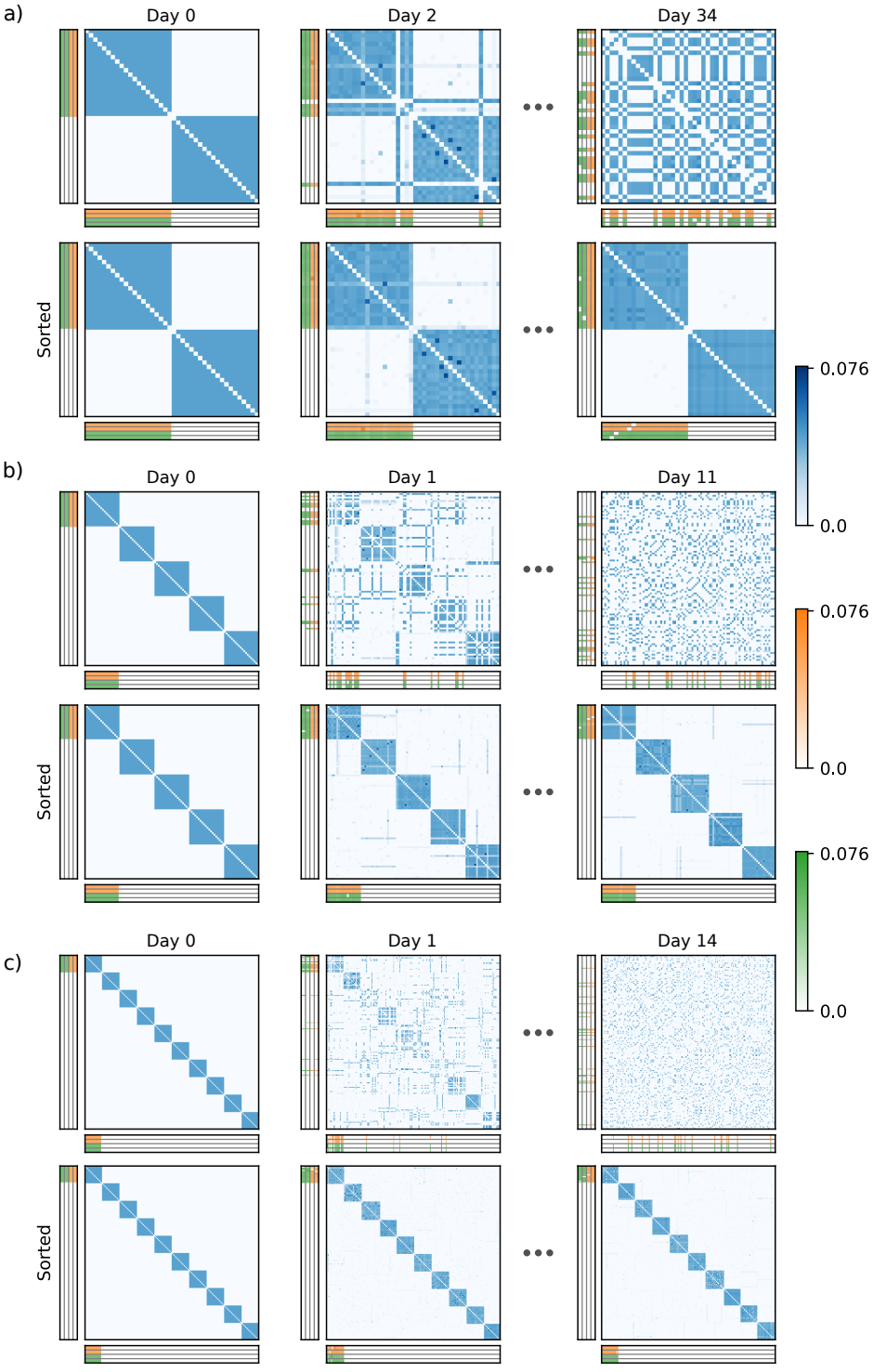

Figure B.15: Networks of binary neurons with two (a), five (b) and ten (c) drifting assemblies. Display is like in main text Fig. 2. First column: Initial state of networks. Second column, after two or one day: several interior neurons switched to a new assembly. Third column, after one complete remodeling: The assemblies have drifted away, the weight matrix has also completely remodeled. We note that networks with different numbers of assemblies have different numbers of neurons and different spike thresholds of the inhibitory unit.

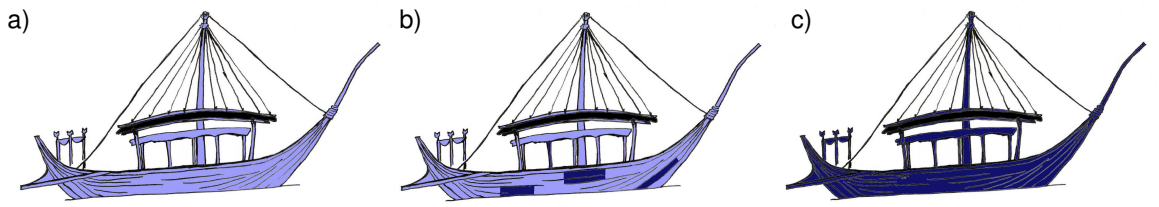

Figure B.16: Theseus' ship. (a) The ship after Theseus' return: all parts are original (light blue). (b) With time, some parts needed to be replaced (dark blue). (c) After a long time, all parts have been replaced; there are no original parts left.

B.2 Family resemblance and identity

L. Wittgenstein uses the analogy of a thread to clarify the relation of the different objects denoted by a word. They need not be directly related by overarching commonalities, but often possess many partial, overlapping "family resemblances", as explained in paragraphs 66 and 67 of ref. [3]:

"66. Consider, for example, the activities that we call "games". I mean board-games, card-games, ball-games, athletic games, and so on. What is common to them all? – Don't say: "They must have something in common, or they would not be called 'games' " – but look and see whether there is anything common to all. – For if you look at them, you won't see something that is common to all, but similarities, affinities, and a whole series of them at that. (...)

67. I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family – build, features, colour of eyes, gait, temperament, and so on and so forth – overlap and criss-cross in the same way. – And I shall say: 'games' form a family. And likewise the kinds of number, for example, form a family. Why do we call something a "number"? Well, perhaps because it has a – direct – affinity with several things that have hitherto been called "number"; and this can be said to give it an indirect affinity with other things that we also call "numbers". And we extend our concept of number, as in spinning a thread we twist fibre on fibre. And the strength of the thread resides not in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres."

Different parts consisting of different ensembles of fibers are parts of the same thread. The overlaps of the fibers give rise to identity over space. Likewise in our model the temporal overlaps of neurons participating in the different assembly realizations give rise to the identity of the memory over time. The various family resemblances are in our model the similarities of neuron ensembles forming the assemblies at close-by times. We note that the thread analogy is sufficiently flexible to cover the change of memories by learning new facts, for example about apples: If a thread becomes larger, of a different material or loosely intertwined with other threads, it still maintains its identity.

The question of identity was already a central one in ancient Greek philosophy. Heraclitus of Ephesus saw objects in continuous change over time. Well known are aphorisms that shall paraphrase his ideas such as "It is not possible to step twice into the same river" [4, 5], with a continuation: "or to come into contact twice with a mortal being in the same state". Another famous aphorism of this kind is cited at the end of our main text.

The ancient historian Plutarch recounts a puzzle on identity, the "Ship of Theseus": After its famous trip this ship is displayed in Athens. Over time one part after the other is gradually replaced. At some point no piece is original anymore, see Fig. B.16. Is it nevertheless Theseus' ship? T. Hobbes added a twist, suggesting that the original parts might be collected somewhere upon their replacement. If in the end they are put together again to a ship, which one is now Theseus' ship [6]? This is another analogy to the assembly model that we propose. The question of identity is for our assemblies solved by the input and output neurons: They connect to the neuron ensemble that forms the memory representation. Suppose that after some time all neurons that originally formed, for example, the apple assembly have been exchanged. This assembly now consists of different neurons (like the displayed ship of different parts), but it is the apple assembly, since the right input and output neurons connect to it; the new neuron ensemble forming the assembly mediates the correct behavior. Now assume that the neurons that originally formed the apple assembly connect later, for some reason, to an assembly again (the original ship parts are put together again). This assembly will not be the apple assembly, since suitable input and output neurons are not connected to it.

B.3 Associative memory property and input-output functionality of assemblies

This section checks that assemblies of LIF neurons have the associative memory property and that their circuits of inputs, assemblies and outputs have basic input-output functionality. For the former, we suppress the periphery neuron activity during part of the network simulation, Fig. B.2a. We then still observe assembly reactivation on a background of sparse activity. Since the sparse background activity will only activate part of an assembly, this shows that partial activation can lead to recall (complete or near complete assembly reactivation) as required for associative memory. In presence of periphery neurons, spontaneous assembly reactivation is more frequent, Fig. B.2b. This is expected because the periphery neurons contribute background spikes and help amplifying assembly activity. Further our circuits are functional in the sense that sufficiently strong stimulation of input neurons activates an assembly, which in turn activates its output neurons, Fig. B.2c. Activation is specific, other assemblies and periphery neurons do not increase their activity. All simulations in Fig. B.2 are done after the first complete remodeling of the system was detected. We set this time to $t = 0$s. In Fig. B.2a,b the weight matrices are kept constant at their values at $t = 0$s. This prevents compensation of the missing periphery neurons in Fig. B.2a. Fig. B.2b shows the beginning of the simulation analyzed in main text Fig. 8a. In Fig. B.2c external stimulation forces the input neurons to be highly active for 0.5s after $t = 4.75$s. The resulting spiking activity does not destroy the circuit structure. All of these observations also hold for the LIF model where spontaneous synaptic turnover drives the assembly drift (Fig. B.12). In this case, not only the weight matrices but also the connectivity matrices are kept constant in Fig. B.12a,b and in Fig. B.12c the external stimulation is applied for 1s after $t = 4.5$s.

B.4 Spontaneous synaptic turnover drives assembly drift in Fig. 5

If the incomplete synaptic connectivity of the network in main text Fig. 5 is kept constant, the assemblies do not drift, see Fig. B.10. There is an initial phase where the assemblies adapt to the connectivity matrix: neurons that receive few input connections from their assembly leave it. In the simulation of Fig. B.10, one neuron changes assembly at around one hour simulated time, see also Fig. B.11. No more changes occur until the end of the simulation at 100 hours.

B.5 Random walk models based on statistics of weight changes

Simulations of a model network or a random walk model allow us to measure the mean and the standard deviations of the weight change Δw_1 of the summed input weight w_1 a neuron receives from assembly 1 for a given value of this weight (Fig. 4). In this section we initially construct a Markovian random walk model for the dynamics of w_1 based on these quantities. We then approximate it by a diffusion process. The obtained models allow us to analytically and numerically compute stationary probability densities and thereby detect metastable states. By removing the drift or homogenizing the noise, we can selectively study the impact of the fluctuations or the average of the weight updates on the switching dynamics.

We use the sampled mean $\overline{\Delta w_1}(w_1)$ and standard deviation $\mathrm{Std}(\Delta w_1)(w_1)$ of weight changes to construct a first random walk model for neuron switching between assemblies. Since the sampling interval is sufficiently long, we can assume that the change in w_1 depends only on its previous

value (Markov assumption). For simplicity we further assume that noise is normally distributed. We therefore simulate the weight dynamics as $w_1(t+1) = w_1(t) + \overline{\Delta w_1}(w_1(t)) + \xi(w_1(t))$, where $\xi(w_1(t)) \sim \mathcal{N}(0, \mathrm{Std}(\Delta w_1)(w_1(t)))$, and w_1 is clipped to the interval $[0, 1]$ after each step. We find that the stationary probability density functions of w_1 for the full models, $p_{\mathrm{full}}(w_1)$, and those of the corresponding Markov simulations are in acceptable agreement, Figs. B.7a, B.8e, B.9e. Deviations are likely due to the non-Gaussianity of the weight change distributions in the network simulations. To study the contribution of weight update fluctuations, we repeat the Markov simulations, but without including the effect of the mean, $w_1(t+1) = w_1(t)+\xi(w_1(t))$. We find that this process also agrees well with the full models, Fig. B.7b, B.8f, B.9f. Finally, to examine the contribution of mean weight change, we calculate the average noise standard deviation, $\overline{\mathrm{Std}(\Delta w_1)(w_1)} = \int_0^1 \mathrm{Std}(\Delta w_1)(w_1) p_{\mathrm{full}}(w_1) dw_1$, and simulate the dynamics with state-independent noise $\xi \sim \mathcal{N}(0, \overline{\mathrm{Std}(\Delta w_1)(w_1(t))})$. With this noise, the neuron does not leave its potential valleys (cf. main text Fig. 5c) within the simulated periods: for the LIF model, the neuron stays with the assemblies, for the binary model it stays with the assemblies or gets trapped in the middle between them. Thus, also for homogenized noise the dynamics show crucial aspects of the full dynamics. We conclude that both drift and noise inhomogeneity contribute to the switching dynamics. The results further indicate that the inhomogeneous noise is already sufficient to generate the crucial features of the stationary distribution, the meta-stable states and the switching. Noise-induced multistability has been observed in different models and natural systems before, e.g. for electrical and chemical oscillations, populations dynamics and foraging behavior [7–10].

We can analytically obtain the stationary probability density functions for the Markov simulations using a diffusion approximation. For this we interpret the simulated weight dynamics as the Euler-Maruyama discretization of an Itô stochastic differential equation [11]. With a time step equal to one we can directly read off the coefficients specifying the drift and the noise strength in the stochastic differential equation: they equal the mean $\overline{\Delta w_1}(w_1)$ and the standard deviation $\mathrm{Std}(\Delta w_1)(w_1)$ of the weight updates. In particular they both depend on the weight w_1. The normalized stationary solution of the Fokker-Planck equation associated to the stochastic differential equation is given by

$$p_{\mathrm{FP}}(w_1) = \frac{\mathcal{N}}{\mathrm{Std}(\Delta w_1)^2(w_1)} \exp\left[2 \int_0^{w_1} du \, \frac{\overline{\Delta w_1}(u)}{\mathrm{Std}(\Delta w_1)^2(u)}\right] \tag{B.1}$$

for $w_1 \in [0, 1]$ and with a normalization constant $\mathcal{N}$ such that $\int_0^1 dw_1 \, p_{\mathrm{FP}}(w_1) = 1$. Here we use reflecting boundary conditions to keep the probability density in the interval $[0, 1]$. For the modified Markov simulations the expression holds with drift or diffusion coefficient replaced accordingly. The resulting stationary probability densities agree with the Markov simulations except at the boundaries, Fig. B.7 (in (c) up to regional scaling factors). The deviations at the boundaries originate from the clipping of w_1 in the Markov simulations; if we replace the clipping by reflecting the increments at the boundaries, the densities also agree there.

B.6 Random walk model from first principles for the switching dynamics in LIF networks

In this section we derive an effective random walk model from first principles, for the neuron switching dynamics in LIF networks where noisy autonomous activity drives the assembly drift (main text

Fig. 2). In the random walk model, a single "test neuron" spikes at a fixed background rate and with an input weight dependent probability when its current or another assembly reactivates. The different assemblies reactivate randomly at a fixed rate and their neurons generate sparse background activity as well. The resulting spike timing dependent plasticity and homeostatic normalization lead to a random walk of the neuron's summed input weights from the different assemblies, which determine its assembly membership and its probability to be activated by a reactivating assembly. Model parameters like the background spike rate and the reactivation rate of the assemblies are chosen to agree with the LIF network simulations main text Fig. 4, to render the random walk model comparable to the network simulation (Supplementary Note 7).

We focus on the test neuron's summed input weights $w_a \geq 0$ from the different assemblies $a = 1, \ldots, n_{\text{asbly}}$, where n_{asbly} is the number of assemblies in the network. The small effect of the neuron on the assemblies and their activities is neglected. Homeostatic plasticity normalizes the sum of the w_a to w_{sum}, which we set to one, thus $\sum_a w_a = 1$. We further use that $w_{\text{sum}}/N_a \leq w_{\text{max}}$ for the sizes N_a of the assemblies. Thus, on the level of the summed input weights from the assemblies, the constraint set by homeostatic normalization is stronger than the upper bound set by the individual maximal synaptic weight. For simplicity, we use a fixed and identical size N_{asbly} for all assemblies. From the perspective of the neuron, the assemblies are regarded as static except for the plasticity of the summed input weights w_a from them. In our LIF networks, weight changes are driven by STDP combined with homeostatic normalization. Pairs of spikes of the test neuron and an assembly neuron contribute to changes of the summed synaptic input weight w_a from the assembly to the test neuron according to the STDP rule (see Methods). Spontaneous assembly reactivations occur at rates r_a for the different assemblies. Again for simplicity, we use an identical reactivation rate r_{asbly} for all assemblies and measure it in our simulations. The activity of the test neuron consists of irregular background spiking, which is modeled as an independent Poisson process with rate r_{bg}. Furthermore, the neuron can be coactivated by an assembly reactivation with a probability $p_{\text{coact}}(w_a)$ depending on the momentary input weight w_a from the assembly. In addition to the spontaneous assembly reactivations, the assembly neurons exhibit asynchronous, irregular background spiking as well. As for the test neuron, we model this background spiking activity by independent Poisson processes with identical rates r_{bg}.

To specify our random walk model we estimate the weight changes induced by STDP around a single assembly reactivation. Motivated by our network simulations (see manuscript Fig. 8c), we assume that each assembly neuron spikes exactly once during a single reactivation and that these spikes occur independently of each other with a given temporal distribution $\rho_{\text{react}}(t)$. The assembly reactivations in our LIF networks are synchronous with a temporal duration of at most 15 ms, and, hence, one may consider approximating the temporal distribution by a Dirac delta function. We nevertheless choose to include the finite temporal duration here as it has a considerable effect on the estimated weight changes and on the coactivation probability $p_{\text{coact}}(w_a)$. For the explicit formulas below we coarsely approximate this temporal distribution $\rho_{\text{react}}(t)$ by a uniform density of duration τ_{react}. We write the STDP window as $\Delta w_{ij}(\Delta t) = \eta\, h(\Delta t)$, where the amplitude η has the units of the weights and, thus, is given as a fraction of w_{sum}. The dimensionless function $h(\Delta t)$ describes the shape of the STDP window, see Fig. B.1a. By separating the spiking activity into reactivation and background spikes of the assembly neurons or of the test neuron, we obtain the following four contributions.

First, the reactivation of, say, assembly a can coincide to some degree with the random background spiking of the test neuron. The corresponding STDP-induced weight change due to a single reactivation

event takes the form

$$\Delta w_a^{\mathrm{bg,react}} = \eta \sum_{t^a_{\mathrm{react}}} \sum_{t_{\mathrm{bg}}} h\left(t_{\mathrm{bg}} - t^a_{\mathrm{react}}\right) , \tag{B.2}$$

where t_{bg} indicate the test neuron's background spike times, which follow an independent Poisson process with rate r_{bg}, while t^a_{react} denote the (by assumption exactly N_a) spike times of the neurons in the assembly. This contribution to the weight change is a random quantity because the spike times are distributed randomly. Its mean and standard deviation under the given assumptions can be computed by taking averages over the different sets of spike times. We start with the mean:

$$\begin{aligned}
\overline{\Delta w_a^{\mathrm{bg,react}}} &= \eta \left\langle \sum_{t^a_{\mathrm{react}}} \sum_{t_{\mathrm{bg}}} h\left(t_{\mathrm{bg}} - t^a_{\mathrm{react}}\right) \right\rangle_{t_{\mathrm{bg}}, t^a_{\mathrm{react}}} \\
&= \eta\, N_a \int_{-\infty}^{+\infty} \left(r_{\mathrm{bg}} \int_{-\infty}^{+\infty} h(s-t)\, ds \right) \rho_{\mathrm{react}}(t)\, dt \\
&= \eta\, N_a\, r_{\mathrm{bg}} \int_{-\infty}^{+\infty} h(\Delta t)\, d\Delta t \,,
\end{aligned} \tag{B.3}$$

where we use that the average of the summed values of a function $f(t)$ at the random points t_{Poi} of a homogeneous Poisson point process with rate r_{Poi} is given by

$$\left\langle \sum_{t_{\mathrm{Poi}}} f\left(t_{\mathrm{Poi}}\right) \right\rangle_{t_{\mathrm{Poi}}} = r_{\mathrm{Poi}} \int_{-\infty}^{\infty} f(s)\, ds \tag{B.4}$$

(Campbell's theorem for the mean, ref. [12]). We note that the mean, Eq. (B.3), does not depend on the temporal distribution $\rho_{\mathrm{react}}(t)$ of the assembly reactivation spikes. To obtain the variance of the weight change contribution we use the law of total variance,

$$\mathrm{Var}\left(\Delta w_a^{\mathrm{bg,react}}\right) = \left\langle \mathrm{Var}\left(\Delta w_a^{\mathrm{bg,react}} \middle| t^a_{\mathrm{react}}\right) \right\rangle_{t^a_{\mathrm{react}}} + \mathrm{Var}\left(\mathrm{E}\left(\Delta w_a^{\mathrm{bg,react}} \middle| t^a_{\mathrm{react}}\right) \right)_{t^a_{\mathrm{react}}} , \tag{B.5}$$

where $\mathrm{E}\left(\Delta w_a^{\mathrm{bg,react}} \middle| t^a_{\mathrm{react}}\right)$ is the conditional mean (or expectation value). We first compute the conditional variance given the reactivation spike times:

$$\begin{aligned}
\mathrm{Var}\left(\Delta w_a^{\mathrm{bg,react}} \middle| t^a_{\mathrm{react}}\right) &= \eta^2\, \mathrm{Var}\left(\sum_{t_{\mathrm{bg}}} \sum_{t^a_{\mathrm{react}}} h\left(t_{\mathrm{bg}} - t^a_{\mathrm{react}}\right) \middle| t^a_{\mathrm{react}} \right) \\
&= \eta^2 r_{\mathrm{bg}} \int_{-\infty}^{+\infty} \left(\sum_{t^a_{\mathrm{react}}} h\left(s - t^a_{\mathrm{react}}\right) \right)^2 ds \,,
\end{aligned} \tag{B.6}$$

where we use that for a homogeneous Poisson process with the notation of Eq. (B.4) the variance of

the random sum reads

$$\mathrm{Var}\left(\sum_{t_{\mathrm{Poi}}} f\left(t_{\mathrm{Poi}}\right)\right) = r_{\mathrm{Poi}} \int_{-\infty}^{\infty} \left(f(s)\right)^2 \, ds \tag{B.7}$$

(Campbell's theorem for the variance, ref. [12]). The conditional mean given the reactivation spike times equals the unconditional mean,

$$\begin{aligned}
\mathrm{E}\left(\Delta w_a^{\mathrm{bg,react}} \middle| t_{\mathrm{react}}^a\right) &= \eta \sum_{t_{\mathrm{react}}^a} \left\langle \sum_{t_{\mathrm{bg}}} h\left(t_{\mathrm{bg}} - t_{\mathrm{react}}^a\right) \right\rangle_{t_{\mathrm{bg}}} \\
&= \eta \, r_{\mathrm{bg}} \sum_{t_{\mathrm{react}}^a} \int_{-\infty}^{+\infty} h\left(s - t_{\mathrm{react}}^a\right) \, ds \\
&= \eta \, N_a \, r_{\mathrm{bg}} \int_{-\infty}^{+\infty} h(\Delta t) \, d\Delta t \,,
\end{aligned} \tag{B.8}$$

where we used the substitution $\Delta t = s - t_{\mathrm{react}}^a$ in the integral of the second line. The conditional mean's independence of t_{react} implies that its variance $\mathrm{Var}\left(\mathrm{E}\left(\Delta w_a^{\mathrm{bg,react}} \middle| t_{\mathrm{react}}^a\right)\right)_{t_{\mathrm{react}}^a}$ vanishes. Eq. (B.5) then yields the unconditional variance

$$\begin{aligned}
\mathrm{Var}\left(\Delta w_a^{\mathrm{bg,react}}\right) &= \left\langle \mathrm{Var}\left(\Delta w_a^{\mathrm{bg,react}} \middle| t_{\mathrm{react}}^a\right) \right\rangle_{t_{\mathrm{react}}^a} \\
&= \eta^2 r_{\mathrm{bg}} \int_{-\infty}^{+\infty} \left\langle \left(\sum_{t_{\mathrm{react}}^a} h\left(s - t_{\mathrm{react}}^a\right) \right)^2 \right\rangle_{t_{\mathrm{react}}^a} ds \\
&= \eta^2 r_{\mathrm{bg}} \left[N_a (N_a - 1) \int_{-\infty}^{+\infty} \tilde{h}^2(s) \, ds + N_a \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h^2(s-t) \, \rho_{\mathrm{react}}(t) \, dt \, ds \right] \\
&= \eta^2 r_{\mathrm{bg}} \left[N_a^2 \int_{-\infty}^{+\infty} \tilde{h}^2(s) \, ds + N_a \left(\int_{-\infty}^{+\infty} h^2(s) \, ds - \int_{-\infty}^{+\infty} \tilde{h}^2(s) \, ds \right) \right],
\end{aligned} \tag{B.9}$$

where $\tilde{h}(s) = \int_{-\infty}^{+\infty} h(s-t) \rho_{\mathrm{react}}(t) \, dt$ is the convolution of the STDP window with the reactivation spike density. To compute the average over the reactivation spike times t_{react}^a in the second line of Eq. (B.9) we expanded the square and used our assumption that these times are independently and identically distributed with the density $\rho_{\mathrm{react}}(t)$. For the parameter regime of our LIF networks we can neglect the term proportional to N_a in the last line of Eq. (B.9) and obtain a compact expression for the standard deviation

$$\mathrm{Std}\left(\Delta w_a^{\mathrm{bg,react}}\right) \approx \eta \, N_a \left(r_{\mathrm{bg}} \int_{-\infty}^{+\infty} \tilde{h}^2(s) \, ds \right)^{1/2} . \tag{B.10}$$

For the STDP window function employed in the LIF network models and a uniform reactivation spike density, the integrals in Eqs. (B.3) and (B.10) can be computed analytically.

The second contribution to the STDP-induced weight change arises because an assembly can coactivate the test neuron during a reactivation event. This happens in particular if the test neuron is

part of the assembly. We describe our model for the coactivation probability $p_{\text{coact}}(w_a)$ below. We assume that the coactivated neuron independently emits one spike with the same temporal distribution $\rho_{\text{react}}(t)$ as the (other) assembly reactivation spikes such that the resulting weight change can be written as

$$\Delta w_a^{\text{coact}} = \eta \sum_{t_{\text{react}}^a} h\left(t_{\text{coact}} - t_{\text{react}}^a\right) \ . \tag{B.11}$$

Here t_{coact} is the coactivation spike time while t_{react}^a are the N_a assembly reactivation spike times as before. Under our assumptions the differences $t_{\text{coact}} - t_{\text{react}}^a$ are identically distributed with probability density $\hat{\rho}(s) = \int_{-\infty}^{+\infty} \rho_{\text{react}}(s+u)\rho_{\text{react}}(u)\, du$, which results in a symmetric triangular (or "hat") density $\max(\tau_{\text{react}} - |s|, 0)/\tau_{\text{react}}^2$ with one-sided temporal width τ_{react} for the uniform approximation of the spike density. Therefore, the mean of this weight change contribution is given by

$$\begin{aligned}
\overline{\Delta w_a^{\text{coact}}} &= \eta \left\langle \sum_{t_{\text{react}}^a} h\left(t_{\text{coact}} - t_{\text{react}}^a\right) \right\rangle_{t_{\text{coact}}, t_{\text{react}}^a} \\
&= \eta\, N_a \int_{-\infty}^{+\infty} h(s)\hat{\rho}(s)\, ds \\
&\approx \eta\, N_a \left(1 - \frac{1}{3}\frac{\tau_{\text{react}}}{\tau_{\text{pot}}}\right) ,
\end{aligned} \tag{B.12}$$

where the third line is a useful approximation appropriate for $\tau_{\text{react}} \leq \tau_{\text{pot}}$; it follows from approximating the central, positive part of the STDP window by another symmetric triangular function $\eta\left(1 - |\Delta t|/\tau_{\text{pot}}\right)$ with peak value η and zero-crossings at $\pm\tau_{\text{pot}}$. The condition $\tau_{\text{react}} \leq \tau_{\text{pot}}$ also implies that this weight change consists only of potentiation (LTP). We present the approximation here because it compactly exposes the effect of the temporal distribution of the reactivation spike times on the STDP-induced weight change originating from the test neuron's coactivation with the assembly. For the variance $\text{Var}\left(\Delta w_a^{\text{coact}}\right)$ of this weight change contribution we note that the differences $t_{\text{coact}} - t_{\text{react}}^a$ are not independently distributed since they all contain the same random time t_{coact}. We thus proceed by computing the second moment as follows:

$$
\begin{aligned}
\mathrm{Var}\left(\Delta w_a^{\mathrm{coact}}\right) &= \eta^2 \left\langle \left(\sum_{t^a_{\mathrm{react}}} h\left(t_{\mathrm{coact}} - t^a_{\mathrm{react}}\right) \right)^2 \right\rangle_{t_{\mathrm{coact}}, t^a_{\mathrm{react}}} - \left(\overline{\Delta w_a^{\mathrm{coact}}}\right)^2 \\
&= \eta^2 \int_{-\infty}^{+\infty} \left\langle \left(\sum_{t^a_{\mathrm{react}}} h\left(s - t^a_{\mathrm{react}}\right) \right)^2 \right\rangle_{t^a_{\mathrm{react}}} \rho_{\mathrm{react}}(s)\, ds - \left(\overline{\Delta w_a^{\mathrm{coact}}}\right)^2 \\
&= \eta^2 \left[N_a(N_a - 1) \int_{-\infty}^{+\infty} \tilde{h}^2(s)\, \rho_{\mathrm{react}}(s)\, ds \right. \\
&\qquad \left. + N_a \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h^2\left(s - t\right)\, \rho_{\mathrm{react}}(t)\, \rho_{\mathrm{react}}(s)\, dt\, ds \right] - \left(\overline{\Delta w_a^{\mathrm{coact}}}\right)^2 \\
&= \eta^2 \left[N_a(N_a - 1) \int_{-\infty}^{+\infty} \tilde{h}^2(s)\, \rho_{\mathrm{react}}(s)\, ds \right. \\
&\qquad \left. + N_a \int_{-\infty}^{+\infty} h^2\left(s\right)\, \hat{\rho}(s)\, ds - N_a^2 \left(\int_{-\infty}^{+\infty} h(s) \hat{\rho}(s)\, ds \right)^2 \right], \qquad \text{(B.13)}
\end{aligned}
$$

where $\tilde{h}(s) = \int_{-\infty}^{+\infty} h(s-t)\rho_{\mathrm{react}}(t)\, dt$ is the convolution of the STDP window with the reactivation spike density as in Eq. (B.9). For $\tau_{\mathrm{react}} \leq \tau_{\mathrm{pot}}$ we can obtain a compact formula for the variance using the same approximations as for the mean weight change in Eq. (B.12). To evaluate the first integral in the last line of Eq. (B.13) we only need to know $\tilde{h}(s)$ for values of s where $\rho_{\mathrm{react}}(s)$ is nonzero. For these values of s, we approximate $\tilde{h}(s)$ by convolving the uniform spike density of duration τ_{react} (for concreteness centered at zero) with the central, positive part of the STDP window shape, which we again replace by the symmetric triangular function $\left(1 - |s|/\tau_{\mathrm{pot}}\right)$,

$$
\tilde{h}(s) \approx 1 - \frac{1}{4}\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}} - \frac{s^2}{\tau_{\mathrm{react}}\tau_{\mathrm{pot}}}, \ |s| \leq \frac{\tau_{\mathrm{react}}}{2}. \qquad \text{(B.14)}
$$

For the second integral in the last line of Eq. (B.13) we integrate $\left(1 - |s|/\tau_{\mathrm{pot}}\right)^2$ with respect to the symmetric triangular density $\max(\tau_{\mathrm{react}} - |s|, 0)/\tau^2_{\mathrm{react}}$. Together with the approximation of the mean in Eq. (B.12) we therefore obtain

$$
\begin{aligned}
\mathrm{Var}\left(\Delta w_a^{\mathrm{coact}}\right) &\approx \eta^2 \left[N_a(N_a - 1) \left(1 - \frac{2}{3}\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}} + \frac{7}{60}\left(\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}}\right)^2 \right) + N_a \left(1 - \frac{2}{3}\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}} + \frac{1}{6}\left(\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}}\right)^2 \right) \right. \\
&\qquad \left. - N_a^2 \left(1 - \frac{2}{3}\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}} + \frac{1}{9}\left(\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}}\right)^2 \right) \right] \\
&= \eta^2 \left(\frac{N_a^2}{180} + \frac{N_a}{20} \right) \left(\frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}} \right)^2, \qquad \text{(B.15)}
\end{aligned}
$$

which gives the standard deviation of the weight change

$$\mathrm{Std}\left(\Delta w_a^{\mathrm{coact}}\right) \approx \eta\, N_a \left(\frac{1}{180} + \frac{1}{20N_a}\right)^{1/2} \frac{\tau_{\mathrm{react}}}{\tau_{\mathrm{pot}}}\,. \tag{B.16}$$

We note that the variability of this weight change contribution solely originates from the variability of the spike times within an assembly reactivation and, thus, vanishes if its temporal duration approaches zero, i.e. for $\tau_{\mathrm{react}} \to 0$.

The third and fourth contributions to the STDP-induced weight change arise from the asynchronous irregular background spiking of the assembly neurons, which we model as independent Poisson processes with identical rates r_{bg}. These weight change contributions affect the summed input weights w_b from all assemblies $b = 1, \ldots, n_{\mathrm{asbly}}$ in the network, while the first and the second contributions only affect the summed input weight w_a from the assembly a that has currently reactivated. The third contribution is only present when the test neuron is coactived by an assembly reactivation. In this case it emits a spike, which can coincide to some degree with the background spiking of the assembly neurons. The corresponding STDP-induced changes of the summed input weights w_b from the different assemblies $b = 1, \ldots, n_{\mathrm{asbly}}$ read

$$\Delta w_b^{\mathrm{coact,bg}} = \eta \sum_{t_{\mathrm{bg}}^b} h\left(t_{\mathrm{coact}} - t_{\mathrm{bg}}^b\right)\,, \tag{B.17}$$

where t_{bg}^b indicate the background spike times of all neurons in assembly b while t_{coact} is the coactivation spike time as before. The superposition of independent Poisson processes forms another Poisson process whose rate equals the sum of the individual rates. Therefore, the background spike times t_{bg}^b of all neurons in assembly b follow a Poisson process with rate $N_b\, r_{\mathrm{bg}}$, and we can compute the mean and standard deviation of the third weight change contribution in a similar manner to the first contribution above. Campbell's theorem for the mean, Eq. (B.4), yields

$$\begin{aligned}
\overline{\Delta w_b^{\mathrm{coact,bg}}} &= \eta \left\langle \sum_{t_{\mathrm{bg}}^b} h\left(t_{\mathrm{coact}} - t_{\mathrm{bg}}^b\right) \right\rangle_{t_{\mathrm{coact}}, t_{\mathrm{bg}}^b} \\
&= \eta \int_{-\infty}^{+\infty} \left(N_b\, r_{\mathrm{bg}} \int_{-\infty}^{+\infty} h(t-s)\, ds \right) \rho_{\mathrm{react}}(t)\, dt \\
&= \eta\, N_b\, r_{\mathrm{bg}} \int_{-\infty}^{+\infty} h(\Delta t)\, d\Delta t\,.
\end{aligned} \tag{B.18}$$

To obtain the variance, we employ the law of total variance together with Campbell's theorem for the

variance Eq. (B.7), and use that variance of the conditional mean vanishes as in Eq. (B.9):

$$\begin{aligned}\mathrm{Var}\left(\Delta w_b^{\mathrm{coact,bg}}\right) &= \left\langle \mathrm{Var}\left(\Delta w_b^{\mathrm{coact,bg}}\middle| t_{\mathrm{coact}}\right)\right\rangle_{t_{\mathrm{coact}}} + \mathrm{Var}\left(\mathrm{E}\left(\Delta w_b^{\mathrm{coact,bg}}\middle| t_{\mathrm{coact}}\right)\right)_{t_{\mathrm{coact}}} \\ &= \eta^2 N_b\, r_{\mathrm{bg}} \left\langle \int_{-\infty}^{+\infty} h^2(t_{\mathrm{coact}} - s)\, ds \right\rangle_{t_{\mathrm{coact}}} \\ &= \eta^2 N_b\, r_{\mathrm{bg}} \int_{-\infty}^{+\infty} h^2(\Delta t)\, d\Delta t\,. \end{aligned} \tag{B.19}$$

Hence, in contrast to the standard deviation Eq. (B.10) of the first contribution, the resulting standard deviation

$$\mathrm{Std}\left(\Delta w_b^{\mathrm{coact,bg}}\right) = \eta \left(N_b\, r_{\mathrm{bg}} \int_{-\infty}^{+\infty} h^2(\Delta t)\, d\Delta t \right)^{1/2} \tag{B.20}$$

is proportional to the square root of the number of neurons in the assembly making it significantly smaller than the former.

Finally, also the background spiking of the test neuron can coincide with the background spiking of the assembly neurons, which leads to STDP-induced weight changes that occur independently of the assembly reactivations. Here, we only include the mean of this fourth weight change contribution as it is non-zero for the employed STDP window. We expect that the fluctuations are proportional to the square root of the number of neurons in the assembly as in Eq. (B.20) and are, therefore, small compared to the fluctuations of the first weight change contribution, Eq. (B.10). It is, however, difficult to estimate the standard deviation of this weight change contribution because of the ongoing homeostatic weight normalization. We therefore consider the mean weight change over a not too long time interval without ongoing weight normalization and apply it only it at the end of the interval. As before, we model the background spiking of the test neuron and of the assembly neurons as independent Poisson processes with rates r_{bg} for the test neuron and $N_b\, r_{\mathrm{bg}}$ for assembly b. The average of the resulting (additive) STDP-induced changes of the summed input weights w_b during a time interval of length T is then proportional to the product of the rates [13]:

$$\overline{\Delta w_b^{\mathrm{bg,bg}}}(T) = \eta\, N_b\, r_{\mathrm{bg}}^2\, T \int_{-\infty}^{+\infty} h(\Delta t)\, d\Delta t\,, \tag{B.21}$$

where T should be at least several times longer than the temporal extent of the STDP window $h(\Delta t)$. On the other hand, to approximately implement the ongoing weight normalization the length T should not be too long. In our random walk model we use the mean weight change Eq. (B.21) for the time intervals between successive assembly reactivations. Their average length is $1/\left(n_{\mathrm{asbly}} r_{\mathrm{asbly}}\right)$.

The test neuron's probability $p_{\mathrm{coact}}(w_a)$ of being coactivated by the synchronous assembly reactivation depends on the momentary input weight w_a from the assembly. Because the second weight change contribution from above is only present when the test neuron is coactive with the assembly and also only affects, specifically strengthens, the summed input weight from this assembly, the coactivation probability implicitly gives rise to a weight-dependence of the STDP-induced weight change. It is thus an important part of the random walk model. The probability that a noisy LIF neuron emits a spike in response to synchronous spiking of excitatory inputs has been previously investigated, see e.g. [14, 15]. As the assembly reactivations in our LIF networks consist of synchronous spiking

with approximately one spike per assembly neuron, the response probability is well described by the probability of finding the neuron's fluctuating membrane potential V in the interval $[V_\theta - \gamma\, w_a, V_\theta]$, where V_θ is the spike threshold and $\gamma\, w_a$ is the expected peak value of the compound post-synaptic potential generated by the synchronous inputs. The factor γ thus translates the dimensionless input weight w_a we use in our random walk model to synaptic input weights in terms of their peak EPSPs, as we display them in main text Fig. 2. It also includes an attenuation of the peak value due to the finite temporal duration of the reactivation spike density $\rho_{\text{react}}(t)$; for the uniform density of duration τ_{react}, we have $\gamma = \gamma(\tau_{\text{react}})$. To obtain an explicit formula for the coactivation probability we approximate the membrane potential distribution $p_V(v)$ of the neuron before a reactivation event by a Gaussian density with mean μ_V and standard deviation σ_V [14, 15] leading to

$$\begin{aligned} p_{\text{coact}}(w_a) &= \int_{V_\theta - \gamma\, w_a}^{V_\theta} p_V(v)\, dv \\ &= \frac{1}{2}\left[\text{erf}\left(\frac{\gamma\, w_a}{\sqrt{2}\sigma_V} - \frac{V_\theta - \mu_V}{\sqrt{2}\sigma_V}\right) + \text{erf}\left(\frac{V_\theta - \mu_V}{\sqrt{2}\sigma_V}\right)\right], \end{aligned} \tag{B.22}$$

which has the natural property $p_{\text{coact}}(0) = 0$ and increases in a sigmoidal manner to values close to 1 for $\gamma w_a - (V_\theta - \mu_V) \gg \sqrt{2}\sigma_V$. For simplicity, we use $\mu_V = V_{\text{rest}}$ and $\sigma_V = \sigma$, the mean and the standard deviation of the stationary membrane potential distribution of our LIF model neurons in absence of a threshold and synaptic inputs from other modeled neurons (Methods). We expect that the asynchronous background spiking at low rates of the other neurons does not significantly affects the membrane potential distribution. Alternatively, instead of the Gaussian density in Eq. (B.22) one could use the known stationary density of the LIF neuron with white noise input, see e.g. [16]. Here, we opt for the more explicit approximation.

The obtained STDP-induced weight change contributions around a single assembly reactivation together with the test neuron's coactivation probability enable us to specify our random walk model describing the dynamics of the test neuron's summed input weights w_a from the assemblies. To model the dynamics in continuous time we assume that the times of the spontaneous assembly reactivations follow independent Poisson processes with identical reactivation rates $r_a = r_{\text{asbly}}$. In the random walk model we move along the reactivation events. At the reactivation of an assembly we apply the different weight change contributions as follows. We start with the fourth contribution resulting from the background spiking of both the test neuron and the assembly neurons. All the weights change according to Eq. (B.21) with T given by the time interval since the previous assembly reactivation, which on average equals $1/\left(n_{\text{asbly}} r_{\text{asbly}}\right)$. After the application of the STDP-induced additive changes to the weights, we clip them to $w_a \geq 0$ and employ homeostatic divisive normalization by updating the weights to

$$w_a = \frac{w_a'}{\sum_{b=1}^{n_{\text{asbly}}} w_b'}, \tag{B.23}$$

where w_a' are the changed weights after clipping. We use this updated weight w_a, i.e. including clipping and normalization after application of Eq. (B.21), in the coactivation probability $p_{\text{coact}}(w_a)$ to randomly decide if the test neuron is coactived by the assembly reactivation. Next, we apply the first weight change contribution, which originates from the coincidence of the assembly reactivation with

the background spiking of the test neuron. To include the important fluctuations of this contribution we randomly draw the (additive) weight change from a normal distribution with mean Eq. (B.3) and standard deviation Eq. (B.10). Afterwards, we again clip the changed weights to $w_a \geq 0$ and then normalize them according to Eq. (B.23). In case the test neuron is coactive with the assembly, the neuron's coactivation spike leads to further STDP-induced weight changes, which are characterized above as the second and third contributions. We also randomly draw them from normal distributions with corresponding means and standard deviations, Eqs. (B.12) and (B.16) for the second and Eqs. (B.18) and (B.20) for the third contribution. Of particular importance is the second contribution because it describes how coactivation reinforces the weight from the reactivating assembly. As before, the changed weights are clipped and normalized, where we employ Eq. (B.3) only after the application of the second and third contribution together.After all weight updates we move forward to the next time of an assembly reactivation, which in our model occurs after an exponentially distributed waiting time with mean $1/\left(n_{\text{asbly}} r_{\text{asbly}}\right)$. These steps completely define our effective random walk model of a neuron's input weights from the assemblies in our LIF networks. We note that the random walk model does not incorporate the weight decrease when an assembly reactivates and the neuron stays silent (indirect weight decrease via homeostatic normalization), because this would require the modeling of weight updates within the assembly.

Main text Fig. 4a middle shows a switching event generated by the random walk model and Fig. 4b,c middle displays the mean and standard deviations of the resulting weight updates (including normalization). These statistics describe the weight changes between two consecutive reactivation events. In contrast, the statistics in Fig. 4b,c left of a LIF network (with two assemblies and without periphery neurons) are obtained by sampling the weight changes during average single neuron interspike intervals. In our network simulations, the two used time intervals, however, approximately agree. This is also reflected in the relation $r_{\text{asbly}} + r_{\text{bg}} = n_{\text{asbly}} r_{\text{asbly}}$ of the chosen random walk model parameters (Supplementary Note 7), since the average single neuron rate is approximately $r_{\text{asbly}} + r_{\text{bg}}$ (neurons belonging to an assembly usually reliably spike with it). Therefore, the statistics in Fig. 4b,c left and middle agree well even quantitatively.

B.7 Random walk model from first principles for the switching dynamics in binary networks

In this section we obtain a random walk model for switching in binary networks similar to the one for LIF networks, see previous section. The more abstract nature of the binary neuron model and of the plasticity rule greatly simplifies the derivation. For simplicity, we take into account only those changes to w_1 that occur when the neuron spikes together with reactivation of one of the assemblies; i.e. we neglect the small effects of weight changes between these events and of simultaneous spiking of the neuron with both assemblies. The dynamics of w_1 thus become a discrete-time random walk process with state-dependent noise strength, where time steps with changes in w_1 correspond to the spiking of one of the assemblies together with the neuron. We use time steps of 30 ms, twice the length of those of the binary model, since reactivating assemblies are usually highly active for two consecutive time steps, main text Fig. 8c. We adjust spiking probabilities accordingly. Each assembly is spontaneously active in a time step with probability p_{asbly}. If assembly 1 reactivates and w_1 is larger than the spike threshold θ, the neuron also becomes active. Assembly 2 has the same effect if $w_2 > \theta$. Further, the neuron can be spontaneously active; this happens with probability p_{sp} and enables the

transitions between assemblies. If the neuron and assembly 1 or assembly 2 are active together, the weight from the active assembly is potentiated, by the amount $P(w_1)$ or $P(w_2)$ prior to normalization. The actual change in the weight w_1 including the effect of divisive normalization is

$$\Delta w_1 = \begin{cases} \frac{w_2 P(w_1)}{1+P(w_1)} & \text{if the neuron is coactive with assembly 1,} \\ -\frac{w_1 P(w_2)}{1+P(w_2)} & \text{if the neuron is coactive with assembly 2,} \\ 0 & \text{otherwise.} \end{cases}$$

We set $P(w) = P_{\text{weak}}$, if $w < w_{\text{th}}$ and $P(w) = P_{\text{strong}}$ otherwise, analogously to the binary model. The random walk model has six parameters. The neuron spontaneous spike probability p_{sp}, the spike threshold θ, the threshold w_{th}, and the ratio $P_{\text{weak}}/P_{\text{strong}}$ are set to values matching the binary model. The probability of an assembly spike p_{asbly} is set to that observed in simulations of the binary model. Finally, the plasticity magnitude P_{strong} is chosen such that the switching dynamics resemble those of the binary model.

B.8 Parameters of models used for the simulations

LIF model where noisy autonomous activity drives the drift

Neuron numbers: excitatory neurons: $N_E = 102$; interior neurons: $N_{\text{int}} = 90$; periphery neurons: 12; inhibitory neurons: $N_I = 20$.

Network structure: connection probability between interior neurons: $p_{\text{int}} = 1$; connection probability between interior and periphery neurons: $p_{\text{peri}} = 1$; periphery neurons have no connections between each other; connection probability between excitatory and inhibitory neurons and between inhibitory and inhibitory neurons: 1; there are no self-connections of neurons.

Neuron parameters: spike threshold: $V_\theta = 20\text{mV}$; reset potentia:l $V_0 = 0\text{mV}$; resting potential: $V_{\text{rest}} = 10\text{mV}$; membrane time constant: $\tau_m = 10\text{ms}$; absolute refractory period: $\tau_{\text{ref}} = 5\text{ms}$; sum of input and sum of output weights of an interior neuron: $w_{\text{sum}} = 256.25\text{mV} = \frac{1}{2}\left[\left(N_{\text{asbly}} - 1\right) p_{\text{int}} w_{\text{max}} + N_{\text{peri}} p_{\text{peri}} w_{\text{max,peri}}\right]$, the angular bracketed term is the expected input of an interior neuron from a typical size assembly and its periphery neurons, if all weights were at their individual maximum; sum of input and sum of output weights of a periphery neuron: $w_{\text{sum,peri}} = 225.0\text{mV} = \frac{1}{5}\left[N_{\text{asbly}} p_{\text{peri}} w_{\text{max,peri}}\right]$, the angular bracketed term is the expected input of a periphery neuron from a typical size assembly, if all weights are at their individual maximum; noise input strength: $\sigma = 3.5\text{mV}$.

Excitatory synapses: time constant: $\tau_E = 2\text{ms}$; maximal synaptic strength of synapses between interior neurons: $w_{\text{max}} = 12.5\text{mV}$, evoking a peak EPSP of 1.67mV in a resting postsynaptic neuron; maximal synaptic strength of synapses between interior and periphery neurons: $w_{\text{max,peri}} = 37.5\text{mV}$, evoking a peak EPSP of 5.02mV in a resting postsynaptic neuron; strength of synapses to inhibitory neurons: $w_{\text{E}\to\text{I}} = 5.02\text{mV}$, evoking a peak EPSP of 0.67mV in a resting postsynaptic neuron.

Inhibitory synapses: time constant: $\tau_I = 5\text{ms}$; strength of synapses to excitatory neurons: $w_{\text{I}\to\text{E}} = -5.13\text{mV}$, evoking a peak inhibitory postsynaptic potential (IPSP) of -1.28mV in a resting postsynaptic neuron; strength of synapses to inhibitory neurons: $w_{\text{I}\to\text{I}} = -5.39\text{mV}$ evoking a peak IPSP of -1.35mV in a resting postsynaptic neuron.

STDP window: $\Delta w_{ij}(\Delta t) = \frac{\eta}{a - b(1+\delta)} \left[a \exp(-a\,|\Delta t|) - b(1+\delta)\exp(-b\,|\Delta t|)\right]$, where $\Delta t = t_i - t_j$

is the time difference between the postsynaptic and the presynaptic spike; window amplitude for connections between interior neurons: $\eta = 3.75\text{mV}$; window amplitude for connections between interior and periphery neurons: $\eta = 1.25\text{mV}$; LTP peak at 0ms; LTP decay rate: $a = \frac{1}{\tau_{\text{LTP}}} = \frac{1}{20\text{ms}}$; LTD decay rate: $b = \frac{1}{\tau_{\text{LTD}}} = \frac{1}{40\text{ms}}$; ratio of integrated LTD and LTP: $1 + \delta = 1 + \frac{1}{3}$.

Spontaneous synaptic turnover: none.

Memory representation: number of assemblies: $n_{\text{asbly}} = 3$; initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 30$; periphery neurons per assembly: $N_{\text{peri}} = 4$.

Simulation: time step: 0.25ms; total simulated time: 75 hours.

LIF model for switching mechanism analysis

Same parameters as before with the following exceptions:

Neuron numbers: excitatory neurons: $N_E = 68$; interior neurons: $N_{\text{int}} = 68$; periphery neurons: 0; inhibitory neurons: $N_I = 13$.

Neuron parameters: sum of input and sum of output weights of an interior neuron: $w_{\text{sum}} = 309.375\text{mV} = \frac{3}{4}\left[\left(N_{\text{asbly}} - 1\right) \times\ p_{\text{int}} w_{\text{max}}\right]$.

Excitatory synapses: strength of synapses to inhibitory neurons: $w_{\text{E}\to\text{I}} = 9.10\text{mV}$, evoking a peak EPSP of 1.22mV in a resting postsynaptic neuron.

Inhibitory synapses: strength of synapses to excitatory neurons: $w_{\text{I}\to\text{E}} = -9.52\text{mV}$, evoking a peak inhibitory postsynaptic potential (IPSP) of −2.38mV in a resting postsynaptic neuron; strength of synapses to inhibitory neurons: $w_{\text{I}\to\text{I}} = -10.31\text{mV}$ evoking a peak IPSP of −2.58mV in a resting postsynaptic neuron.

STDP window: window amplitude: $\eta = 5\text{mV}$.

Spontaneous synaptic turnover: none.

Memory representation: number of assemblies: $n_{\text{asbly}} = 2$; initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 34$.

Simulation: total simulated time: 50 hours.

LIF model where spontaneous synaptic turnover drives the drift

Same parameters as for simulations without connectivity remodeling with the following exceptions:

Network structure: connection probability between interior neurons: $p_{\text{int}} = 0.6$; connection probability between interior and periphery neurons: $p_{\text{peri}} = 0.8$;

Neuron parameters: sum of input and sum of output weights of an interior neuron: $w_{\text{sum}} = 253.125\text{mV} = \frac{3}{4}\left[\left(N_{\text{asbly}} - 1\right) \times\ p_{\text{int}} w_{\text{max}} + N_{\text{peri}} p_{\text{peri}} w_{\text{max,peri}}\right]$; sum of input and sum of output weights of a periphery neuron: $w_{\text{sum,peri}} = 225.0\text{mV} = \frac{1}{4}\left[N_{\text{asbly}} p_{\text{peri}} w_{\text{max,peri}}\right]$.

Excitatory synapses: strength of synapses to inhibitory neurons: $w_{\text{E}\to\text{I}} = 4.96\text{mV}$, evoking a peak EPSP of 0.66mV in a resting postsynaptic neuron.

Inhibitory synapses: time constant: $\tau_I = 5\text{ms}$; strength of synapses to excitatory neurons: $w_{\text{I}\to\text{E}} = -5.06\text{mV}$, evoking a peak inhibitory postsynaptic potential (IPSP) of −1.27mV in a resting postsynaptic neuron; strength of synapses to inhibitory neurons: $w_{\text{I}\to\text{I}} = -5.33\text{mV}$ evoking a peak IPSP of −1.33mV in a resting postsynaptic neuron.

STDP window: window amplitude for all connections: $\eta = 1.25\text{mV}$.

Spontaneous synaptic turnover: life and absence time of synapses between interior neurons: $L_{\text{int}} = 2000\text{s}$ and $A_{\text{int}} = 1333.3\text{s}$; life and absence time of synapses between interior and periphery neurons: $L_{\text{peri}} = 2000\text{s}$ and $A_{\text{peri}} = 500\text{s}$.

Simulation: total simulated time: 100 hours.

LIF model where noisy autonomous activity drives the drift, without periphery neurons

Same parameters as for simulations without connectivity remodeling but with periphery neurons with the following exceptions:

Neuron numbers: excitatory neurons: $N_E = 102$; interior neurons: $N_{\text{int}} = 102$; periphery neurons: 0; inhibitory neurons: $N_I = 20$.

Neuron parameters: sum of input and sum of output weights of an interior neuron: $w_{\text{sum}} = 247.5\text{mV} = \frac{3}{5}\left[\left(N_{\text{asbly}} - 1\right) p_{\text{int}} w_{\max}\right]$.

Excitatory synapses: strength of synapses to inhibitory neurons: $w_{\text{E}\to\text{I}} = 4.85\text{mV}$, evoking a peak EPSP of 0.65mV in a resting postsynaptic neuron.

Inhibitory synapses: strength of synapses to excitatory neurons: $w_{\text{I}\to\text{E}} = -4.95\text{mV}$, evoking a peak inhibitory postsynaptic potential (IPSP) of -1.24mV in a resting postsynaptic neuron; strength of synapses to inhibitory neurons: $w_{\text{I}\to\text{I}} = -5.21\text{mV}$ evoking a peak IPSP of -1.30mV in a resting postsynaptic neuron.

STDP window: window amplitude: $\eta = 3.75\text{mV}$.

Memory representation: initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 34$.

Simulation: total simulated time: 50 hours.

LIF random walk model

Same parameters as in the LIF model for the switching mechanism analysis where applicable. Additional parameters:

Assemblies: number of assemblies: $n_{\text{asbly}} = 2$; number of neurons per assembly: $N_{\text{asbly}} = 33$; assembly reactivation rate: $r_{\text{asbly}} = 0.75\text{Hz}$; assembly reactivation duration: $\tau_{\text{react}} = 15\text{ms}$; average waiting time between assembly reactivations: $1/\left(n_{\text{asbly}} r_{\text{asbly}}\right) = 0.66\text{s}$.

Neuron parameters: background spike rate: $r_{\text{bg}} = 0.75\text{Hz}$; coactivation probability: Eq. (B.22) with $\mu_V = V_{\text{rest}} = 10\text{mV}$ and $\sigma_V = \sigma = 3.5\text{mV}$; sum of input weights: $w_{\text{sum}} = 1$; weight translation and attenuation factor: $\gamma = 0.73 \cdot 41.4\text{mV} = 30.2\text{mV}$, equal to the peak compound post-synaptic potential evoked by an assembly reactivation for input weight 1.

STDP window: window shape: $h(\Delta t) = \left[a \exp(-a\,|\Delta t|) - b(1+\delta)\exp(-b\,|\Delta t|)\right] / (a - b(1+\delta))$ with integral $\int_{-\infty}^{+\infty} h(\Delta t)\, d\Delta t = -40\text{ms}$; window amplitude: $\eta = 0.01616$; positive zero-crossing: $\tau_{\text{pot}} = 16.2\text{ms}$

Weight changes: Eqs. (B.3) and (B.10): $\overline{\Delta w_a^{\text{bg,react}}} = -\eta\, N_{\text{asbly}}\, 0.03 = -0.016$ and $\text{Std}\left(\Delta w_a^{\text{bg,react}}\right) = 0.062$; Eqs. (B.12) and (B.16): $\overline{\Delta w_a^{\text{coact}}} = \eta\, N_{\text{asbly}}\, 0.69 = 0.37$ and $\text{Std}\left(\Delta w_a^{\text{coact}}\right) = 0.04$; Eqs. (B.18) and (B.20): $\overline{\Delta w_b^{\text{coact,bg}}} = -\eta\, N_{\text{asbly}}\, 0.03 = -0.016$ and $\text{Std}\left(\Delta w_b^{\text{coact,bg}}\right) = 0.011$; Eq. (B.21): $\overline{\Delta w_b^{\text{bg,bg}}}(T) = -\eta\, N_{\text{asbly}}\, T\, 0.0225\,\text{s}^{-1} = -T\, 0.012\,\text{s}^{-1}$ with time interval length T measured in seconds.

Simulation: event-based; total simulated time: 1000 hrs.

Binary model in Fig. B.6 and in Fig. B.15

Parameters for networks with three assemblies and, in brackets, for networks with two, five and ten assemblies, if different.

Neuron numbers: excitatory neurons: $N_E = 72$ $(48, 120, 240)$; interior neurons: $N_{\text{int}} = 60$ $(40, 100, 200)$; periphery neurons: 12 $(8, 20, 40)$.

Network structure: connection probability between all neurons: $p = 1$; there are no self-connections of neurons.

Neuron parameters: exitatory spike threshold: $\theta = 0.1$; inhibitory spike threshold: $\theta_I = 0.05$ $(0.075, 0.03, 0.015)$; spontaneous spike probability: $p_{\text{sp}} = 0.004$; sum of input and sum of output weights of a neuron: $w_{\text{sum}} = 1$; maximal synaptic strength: $w_{\text{max}} = 0.0769$.

Learning Rule: learning rate for strong synapses: $\eta_{\text{strong}} = 0.03$; learning rate for weak synapses between interior neurons: $\eta_{\text{weak,int}} = 0.075$; learning rate for weak synapses between interior and periphery neurons: $\eta_{\text{weak,peri}} = 0.006$; weak synapse cutoff for synapses between interior neurons: $w_{\text{th,int}} = 0.05 w_{\text{max}}$; weak synapse cutoff for synapses between interior and periphery neurons: $w_{\text{th,peri}} = 0.1 w_{\text{max}}$.

Spontaneous synaptic turnover: none.

Memory representation: number of assemblies: $n_{\text{asbly}} = 3$ $(2, 5, 10)$; initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 20$; periphery neurons per assembly: $N_{\text{peri}} = 4$.

Simulation: time step: 15ms; total simulated time: 121 days.

Binary model for transition mechanism analysis

Neuron numbers: excitatory neurons: $N_E = 72$; interior neurons: $N_{\text{int}} = 72$; periphery neurons: 0.

Network structure: connection probability between all neurons: $p = 1$; there are no self-connections of neurons.

Neuron parameters: exitatory spike threshold: $\theta = 0.1$; inhibitory spike threshold: $\theta_I = 0.1$; spontaneous spike probability: $p_{\text{sp}} = 0.004$; sum of input and sum of output weights of a neuron: $w_{\text{sum}} = 1$; maximal synaptic strength: $w_{\text{max}} = 0.0556$.

Learning Rule: learning rate for strong synapses: $\eta_{\text{strong}} = 0.03$; learning rate for weak synapses between interior neurons: $\eta_{\text{weak,int}} = 0.0045$; weak synapse cutoff for synapses between interior neurons: $w_{\text{th,int}} = 0.05 w_{\text{max}}$.

Spontaneous synaptic turnover: none.

Memory representation: number of assemblies: $n_{\text{asbly}} = 2$; initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 36$.

Simulation: time step: 15ms; total simulated time: 15 days.

Binary model of the fear memory representation

Neuron numbers: excitatory neurons: $N_E = 150$; interior neurons: $N_{\text{int}} = 117$; periphery neurons: 33.

Network structure: connection probability between all neurons: $p = 1$; there are no self-connections of neurons.

Neuron parameters: exitatory spike threshold: $\theta = 0.19$; inhibitory spike threshold: $\theta_I = 0.19$; spontaneous spike probability: $p_{\text{sp}} = 0.05$; sum of input and sum of output weights of a neuron: $w_{\text{sum}} = 1$; maximal synaptic strength: $w_{\text{max}} = 0.04$.

Learning Rule: learning rate for strong synapses: $\eta_{\text{strong}} = 0.01$; learning rate for weak synapses between interior neurons: $\eta_{\text{weak,int}} = 0.0012$; learning rate for weak synapses between interior and periphery neurons: $\eta_{\text{weak,peri}} = 0.0002$; weak synapse cutoff for synapses between interior neurons $w_{\text{th,int}} = 0.05 w_{\max}$; weak synapse cutoff for synapses between interior and periphery neurons: $w_{\text{th,peri}} = 0.1 w_{\max}$.

Spontaneous synaptic turnover: none.

Memory representation: number of assemblies: $n_{\text{asbly}} = 3$; initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 39$; periphery neurons per assembly: $N_{\text{peri}} = 11$.

Simulation: time step: 15ms; total simulated time: 86 days.

Binary model of the XOR gate

Neuron numbers: excitatory neurons: $N_E = 100$; interior neurons: $N_{\text{int}} = 72$; periphery neurons: 28.

Network structure: connection probability between all neurons: $p = 1$; there are no self-connections of neurons.

Neuron parameters: exitatory spike threshold: $\theta = 0.1$; inhibitory spike threshold: $\theta_I = 0.1$; spontaneous spike probability: $p_{\text{sp}} = 0.004$; sum of input and sum of output weights of a neuron: $w_{\text{sum}} = 1$; maximal synaptic strength: $w_{\max} = 0.0455$.

Learning Rule: learning rate for strong synapses: $\eta_{\text{strong}} = 0.03$; learning rate for weak synapses between interior neurons: $\eta_{\text{weak,int}} = 0.0105$; learning rate for weak synapses between interior and periphery neurons $\eta_{\text{weak,peri}} = 0.0009$; weak synapse cutoff for synapses between interior neurons: $w_{\text{th,int}} = 0.05 w_{\max}$; weak synapse cutoff for synapses between interior and periphery neurons: $w_{\text{th,peri}} = 0.15 w_{\max}$.

Spontaneous synaptic turnover: none.

Memory representation: number of assemblies: $n_{\text{asbly}} = 2$; initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 36$; periphery neurons per assembly: $N_{\text{peri}} = 14$.

Simulation: time step: 15ms; total simulated time: 86 days.

Binary random walk model

Neuron parameters: spontaneous spike probability: $p_{\text{sp}} = 0.008$; spike threshold: $\theta = 0.1$.

Learning Rule: plasticity magnitude $P = 0.45$; plasticity magnitude for weak synapses: $P_{\text{weak}} = 0.15P$; weak synapse cutoff: $w_{\text{th}} = 0.05$.

Assembly parameters: spike probability $p_{\text{A}} = 0.0014$.

Simulation: time step: 30ms; total simulated time: 347 days.

Linear Poisson model

Neuron numbers: $N_E = 105$; interior neurons: $N_{\text{int}} = 105$; periphery neurons: 0.

Network structure: connection probability between neurons: $p_{\text{int}} = 0.65$; there are no self-connections of neurons.

Neuron parameters: time constant: $\tau = 10$ms; spontaneous rate: $f_0 = 0.75$Hz; sum of input and sum of output weights of a neuron: $\tau w_{\text{sum}} = 0.25$, maximal synaptic weight: $\tau w_{\max} = 0.0124$.

Plasticity rule: symmetric weight change upon a spike of neuron j: $\Delta w(\Delta f_i) = 0.01/w_{\max} (\Delta f_i - 0.87\text{Hz}) \Delta f_i$, where $\Delta f_i = f_i(t) - f_0$ is the current level of excitation of the post- or presynaptic partner neuron.

Spontaneous synaptic turnover: life and absence time of synapses: $L_{\text{int}} = 6\text{h}$ and $A_{\text{int}} = 3.23\text{h}$.
Memory representation: number of assemblies: $n_{\text{asbly}} = 3$; initial number of interior neurons per assembly: $N_{\text{asbly}}(0) = 35$.
Simulation: event-based; simulated time: 125 days.

References

[1] I. R. Fiete, W. Senn, C. Z. H. Wang and R. H. R. Hahnloser,
Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity., eng, Neuron **65** (2010) 563.

[2] N. Ravid Tannenbaum and Y. Burak,
Shaping Neural Circuits by High Order Synaptic Interactions,
PLoS Computational Biology **12** (2016) 1.

[3] L. Wittgenstein, *Philosophische Untersuchungen/Philosophical investigations*,
ed. by P. M. S. Hacker and J. Schulte, Oxford: Wiley-Blackwell, 2009.

[4] D. W. Graham, "Heraclitus", *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta,
Fall 2019, Metaphysics Research Lab, Stanford University, 2019.

[5] M. E. Rule, T. O'Leary and C. D. Harvey, *Causes and consequences of representational drift*,
Current Opinion in Neurobiology **58** (2019) 141.

[6] R. Wasserman, "Material Constitution", *The Stanford Encyclopedia of Philosophy*,
ed. by E. N. Zalta, Fall 2018, Metaphysics Research Lab, Stanford University, 2018.

[7] L. Arnold, W. Horsthemke and R. Lefever,
White and coloured external noise and transition phenomena in nonlinear systems,
Zeitschrift für Physik B Condensed Matter and Quanta **29** (1978) 367.

[8] W. Horsthemke and R. Lefever, *Noise-Induced Transitions*, Berlin: Springer, 1984.

[9] G. Jetschke, *Mathematik der Selbstorganisation*,
Berlin: VEB Deutscher Verlag der Wissenschaften, 1989.

[10] T. Biancalani, L. Dyson and A. J. McKane,
Noise-Induced Bistable States and Their Mean Switching Time in Foraging Colonies,
Physical Review Letters **112** (2014).

[11] C. Gardiner, *Handbook of Stochastic Methods*, Berlin: Springer, 2002.

[12] J. F. C. Kingman, *Poisson processes*, vol. 3, Oxford Studies in Probability,
Oxford Science Publications, New York: The Clarendon Press Oxford University Press, 1993
viii+104, ISBN: 0-19-853693-3.

[13] R. Kempter, W. Gerstner and J. L. Van Hemmen, *Hebbian learning and spiking neurons*,
Physical Review E **59** (1999) 4498.

[14] S. Goedeke and M. Diesmann,
The mechanism of synchronization in feed-forward neuronal networks,
New Journal of Physics **10** (2008) 015007.

[15] S. Jahnke, R.-M. Memmesheimer and M. Timme,
Propagating synchrony in feed-forward networks,
Frontiers in Computational Neuroscience **7** (2013) 153.

[16] N. Brunel,
Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons,
Journal of Computational Neuroscience **8** (2000) 183.

Acknowledgments

I am very grateful to Raoul-Martin Memmesheimer for far more than mere supervision. A big thanks to my colleagues: Christian Klos, Sven Goedeke, Paul Manz, Paul Züge, Kai Röth, Aditya Gilra, Wilhelm Braun, Slawa Braun, Jonas Nitzsche, Thimo Neugarth, Niels Körner, Talika Beneke, Alexander Schmatz, Bela Erlinghagen, Carla Simon, Fabian Pallasdies, Avleen Sahni, Anna Hellfritzsch, Simon Altrogge, and especially to my family/colleagues: Ivy Chan and Elena Kalle. I am also grateful to Prof. Gerhard von der Emde and Prof. Walter Witke for supporting our experimental endeavors and to Dr. Metsch for advice regarding the thesis.