From dynamics to computations in recurrent neural networks

Francesca Mastrogiuseppe

▶ **To cite this version:**

Francesca Mastrogiuseppe. From dynamics to computations in recurrent neural networks. Physics [physics]. Université Paris sciences et lettres, 2017. English. NNT : 2017PSLEE048 . tel-01820663

HAL Id: tel-01820663
https://theses.hal.science/tel-01820663

Submitted on 22 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'École Normale Supérieure

From dynamics to computations in recurrent neural networks

Ecole doctorale n°564

ÉCOLE DOCTORALE PHYSIQUE EN ÎLE-DE-FRANCE

Spécialité: Physique

Soutenue par
Francesca **MASTROGIUSEPPE**
le 04 décembre 2017

Dirigée par **Vincent HAKIM**
et **Srdjan OSTOJIC**

COMPOSITION DU JURY :

M. LATHAM Peter
Gatsby Computational Neuroscience Unit - UCL London, Rapporteur

Mme. TCHUMATCHENKO Tatjana
Max Planck Institute for Brain Research Frankfurt, Rapporteur

M. HENNEQUIN Guillaume
University of Cambridge, Membre du jury

M. PAKDAMAN Khashayar
Université Paris Diderot, Directeur du jury

M. NADAL Jean-Pierre
École Normale Supérieure, Invité

ENS

PSL RESEARCH UNIVERSITY PARIS

Preface

This thesis collects the work that I have been carrying out in the last three years as Ph.D. student at ENS. During my time here, I collected some results, some failures, and good and bad feedbacks. I have been learning and growing up, and this has been possible thanks to the help of several people that I do need to acknowledge.

To begin with, I would like to deeply thank my two *directeurs de thèse*, who gave me the chance to start this journey. Here I had all the best opportunities to learn: that made me feel an extremely fortunate student. More than anything, I would like to thank Srdjan for having been a generous advisor, a caring supervisor, and – above all – such a nice person to work with. I am extremely thankful to all the members of the jury, who agreed dedicating their time in reading and listening about this research. I would like to thank all the teachers that I met during the summer schools in Woods Hole and Lisbon as well.

I cannot conclude without mentioning all the sweet people that worked and work at the GNT, who contribute to make our office such a friendly and cosy place where to study (and watch movies, and do yoga…). Here I felt at home like a piece of butter on a *tradi*. I would like to thank all the *zie*, and in particular Agnese, for having proofread all my motivational letters and having patiently supported me and my Ph.D. on a daily basis. To conclude, huge thanks to my family and to Riccardo, who remind me that I'm *troppo forte* every time I seem to forget.

September 2017

Contents

Introduction 1

The neural activity from in-vivo cortical recordings displays a large degree of temporal and trial-to-trial irregularity. Multiple layers of variability emerge at multiple time scales and generate complex patterns of cross-correlations among pairs of neurons in the recorded population. Dissecting and characterizing the possible sources of neural variability has been a central research topic in theoretical and computational neuroscience. Ultimately, reconstructing where and how the *noise* is generated could valuably contribute to our more general understanding of how the brain encodes and process information.

One remarkable feature of the brain, that has been suggested to play a major role in generating and shaping variability, is its intricate connectivity structure. Cortical networks, which constitute the fundamental computational units in the mammalian brain, consist of thousands of densely packed neurons that are highly inter-connected through recurrent synapses. Several lines of theoretical research have put forward the hypothesis that irregular activity could emerge in large cortical circuits as a collective dynamical effect. Purely stochastic spiking can indeed be generated within simple and deterministic mathematical models where a large number of elements interact through strong and random recurrent connections. As the latest technological advances allow to appreciate increasingly fine patterns in the complex structure of neural variability, constant theoretical efforts are required to refine and reinvent appropriate network models which can provide a good explanation of the data.

In this chapter, we briefly review the experimental findings that motivate the theoretical studies we propose in this thesis. We examine the most successful modelling results that have been leading our understanding of neural variability across the last two decades. In the last section, we build on those findings to delineate the outline of this dissertation.

1.1 Irregular firing in cortical networks

The mammalian brain is a complex and powerful computing machine. Cortical assemblies of excitatory and inhibitory neurons constitute the hardware which processes the inputs and produces the executable outputs that are demanded in the everyday life tasks. The content of sensory cues, internal judgements and decision variables is encoded in the brain in the form of discrete electrical signals, called action potentials [106, 64]. Action potentials, or simply

spikes, consist of fast depolarizations of the cell membranes.

In order to support highly sophisticated behaviours, the mammalian brain must be capable of extremely reliable and precise computations. The brain language, based on spikes, should thus allow stable and accurate encoding, processing and decoding of information.

The first attempts of systematic in-vivo recordings from cortical cells unveiled a high degree of complexity in the neural code: in most of the experimental setups, the temporal structure of the spike trains seems not to reflect any explicit task-related variable. More surprisingly, the neural code appears to be also strongly variable [93, 115, 43]: the number and the time position of the spikes change dramatically from one trial to the other of the same recording session.

A classical example of in-vivo cortical recording is illustrated in Fig. 1.1. The data refer to the firing activity of a pyramidal cell from the visual area MT of a monkey [120], recorded while the animal is performing a random-dot motion task [21]. Neurons in area MT are known to encode the direction of motion of objects in the visual scene. As a consequence, the cell activity is presumably contributing to the monkey's behavioural response.

The rastergram in Fig. 1.1 **a** shows the firing activity of the cell across several identical repetitions of the task. Across the different trials, the spike trains display stereotyped modulations in the firing rate, which are elicited by a change in the stimulus coherence and luminance. A closer analysis within a smaller time window with almost flat firing rate, instead, reveals a finer temporal scale where the spike occurrence seems to be dominated by randomness (Fig. 1.1 **b**).

Traditionally, the spike train variability has been quantified by looking at the fluctuations in the number of fired action potentials and in the time lag between two consecutive spikes. The inter-spike interval (ISI) histogram measured from in-vivo cortical recordings is typically broad (Fig. 1.1 **c**), and its tail is compatible with an exponential decay [12]. The broadness of the ISI histogram can be measured in terms of the coefficient of variation, given by the ratio of the standard deviation to the mean of the distribution. The value of the coefficient of variation estimated from in-vivo recordings is high, and fluctuates between 0.5 and 1 [126].

The variability in the spike count is quantified instead by comparing the mean number of spikes computed within a fixed time window with the trial-to-trial variance. The values of the mean and the variance are typically found to be comparable (Fig. 1.1 **d**). The curve can be fitted with a slightly super-linear relationship in all the brain areas that have been considered in the literature [43, 144, 22].

Both measures of variability point to a completely stochastic model of spike generation. A simple Poisson model, which assumes that spikes are fired at random from an underlying stationary firing rate, predicts both an exponential ISI histogram with unitary coefficient of variation and a linear increase of the variance with the mean of the spike count.

High levels of variability have been robustly observed across different animal preparations and across several brain areas [75, 56], suggesting that what we perceive as *noise* could be an integrative and fundamental feature of the neural code. Experimental evidence thus quickly turned into fundamental theoretical questions, such as: is there any significance in the occurrence and in the time position of a single spike, or is the information content redundantly stored in the form of average firing rates? Is variability a coding feature, or a constraint that the brain has to deal with?

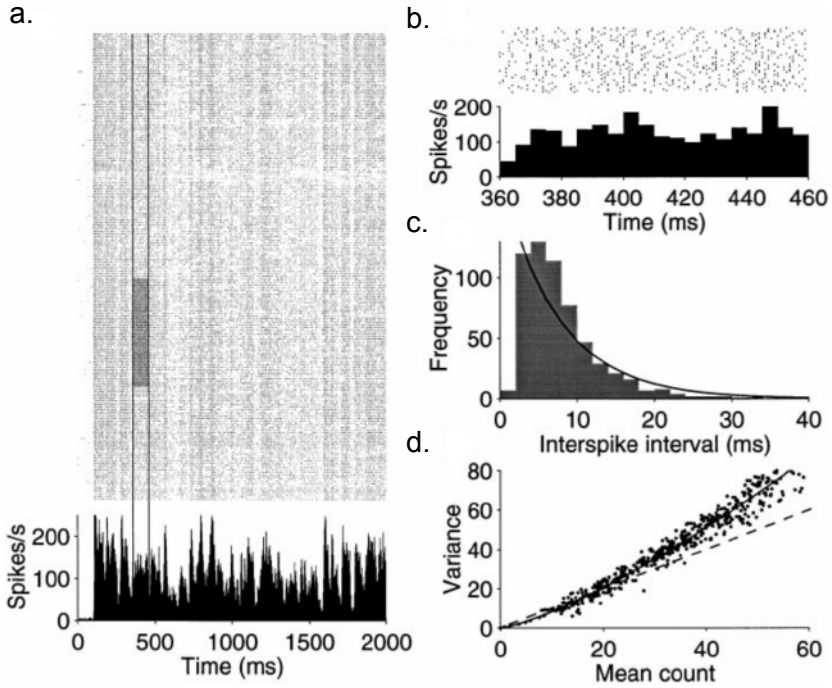


FIGURE 1.1: Variability in a single neuron spiking activity: recordings from area MT of an alert monkey attending the random-dot motion task [21]. **a.** Rastergram and peri-stimulus time histogram. The same identical visual stimulus is presented across 210 different trials. Firing rate modulations are elicited by slow variations in the stimulus coherence and luminance. **b.** Magnified view of the shaded region in **a** (from 360 to 460 ms), where the average firing rate is almost stationary. **c.** Inter-spike interval histogram. The solid line corresponds to the best exponential fit to the data. **d.** Variance in the spike count against the mean spike count. Every dot of the plot is measured in a different time window and from a restricted subset of trials. The best fit polynomial curve is displayed as solid line ($y = x^{1.3}$), while the prediction from a stationary Poisson process is shown as dashed. Adapted from [120].

1.1.1 Irregular inputs, irregular outputs

Although a comprehensive understanding of the cortical code is far from being achieved, major progresses have been made in characterizing the possible mechanicistic sources which underlie the observed variability.

To begin with, neurons are complex and fragile biological devices. As a consequence, a fraction of variability is likely to be generated intrinsically during the input-output process which leads to the spike initiation. Possible sources of stochasticity arise from the finite number of open ion channels in the neuron membrane, from the thermal noise which acts on the charge carriers, or again from the low effectiveness of synaptic transmission [84, 116, 63].

Although such intrinsic sources of noise are likely to contribute, additional experimental evidences from in-vitro setups suggest that their role might be of minor relevance. When they are isolated from their afferent cells, indeed, neurons fairly reliably respond to slowly and fast modulated input currents [82, 27].

On the other hand, the structure of the total input current that neurons receive when they

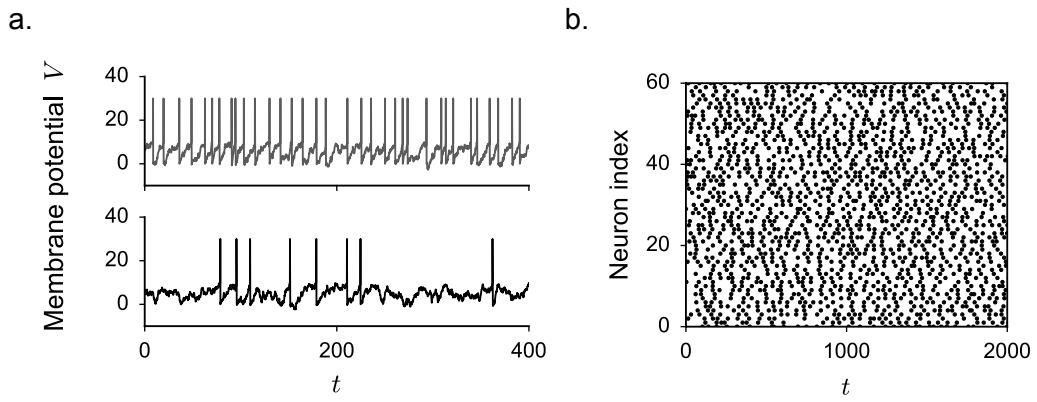


Figure 1.2: Poisson-like firing activity can emerge in model neurons which receive balanced excitatory and inhibitory input currents. **a.** Sample time traces of the membrane potential variable for a leaky integrate-and-fire model [53] of a cortical neuron. The neuron has membrane time constant $\tau = 10$ (all units are arbitrary). The reset is at baseline, and the threshold membrane potential is at $V_{th} = 10$. The neuron receives a stationary excitatory input current together with a stochastic contribution originating from 1000 excitatory and 1000 inhibitory pre-synaptic Poisson spike trains. Top: the mean input value exceeds the threshold potential V_{th}. The membrane potential quickly climbs to the threshold value and thus quite regularly generates spikes. Bottom: the mean input lays slightly below the threshold. In its sub-threshold dynamics, the membrane potential accumulates the Poisson noise, and spikes are emitted as a result of random fluctuations. **b.** Random network models where the strength of excitatory and inhibitory connections is correctly balanced admit a stable state where neurons asynchronously and irregularly emit action potentials. Rasterplot of the simulated spiking activity of 60 units from a larger network of $N = 20000$ leaky integrate-and-fire neurons. Model architecture as in [24].

are integrated in the cortical circuits is largely unknown. It is thus legitimate to hypothesise that neurons do not actively produce, but simply inherit, the noise which is already present at the level of their inputs. Cortical cells, indeed, integrate the action potentials which are generated by several thousands of neighbouring cells. If one assumes that neighbouring neurons fire Poisson spikes, the total input current received by a single cortical neuron is purely stochastic, its mean and variance being determined by the firing rate of its pre-synaptic afferents.

Crucially, a model neuron which integrates an incoming noisy current can operate in a regime where the variability in the input is reflected in its output [142, 53]. In order to obtain irregular outputs, it is critical to assume that the input contributions originating from excitatory and inhibitory pre-synaptic cells loosely balance each other [119, 120].

In those conditions, the mean input is small and elicits solely sub-threshold responses in the post-synaptic membrane potential. Since the variance of the input does not vanish, however, the sub-threshold dynamics is dominated by noise, and spikes can be generated by random fluctuations of the membrane potential (Fig. 1.2 **a.**). In this dynamical regime, the output of the neuron is far from saturation, a desirable feature which allows a wide range of firing rate responses [120].

The output variability can almost completely match the irregularity of the Poisson processes that the neuron receives as input [142]. The output firing rate ϕ_{out} depends on the

statistics of the input current, or equivalently, on the average pre-synaptic firing rate ϕ_{in}. One can write [125]:

$$\phi_{out} = F(\phi_{in}), \tag{1.1}$$

where the function F incorporates both the details of the input connectivity structure and both the single-cell biophysical principles of the spike initiation.

If excitatory and inhibitory currents are correctly balanced, one can finally derive a more global picture where every neuron of the network receives and sends out noise. In a cortical assembly, indeed, every neuron acts both as input and as output. In the hypothesis that different cells can be considered as statistically equivalent objects, a self-consistent network state requires [9, 143, 105]:

$$\phi = F(\phi). \tag{1.2}$$

For a fixed function F, Eq. 1.2 can be used to determine the self-consistent firing rate ϕ at which every network unit is spiking.

Almost twenty years ago, rigorous mathematical analyses have been used to show that this self-consistent solution correspond to a stable collective state for the dynamics of random architectures of excitatory and inhibitory units [24, 145, 11]. In this regime, the input current received by every neuron is dynamically maintained close to the threshold value thanks to the disordered structure of synaptic connections. As a consequence, spikes are driven by fluctuations and different neurons fire asynchronously (Fig. 1.2 **b**). Extremely irregular spike trains emerge, even in absence of external sources of noise, because of the chaotic nature of the high-dimensional attractor underlying the dynamics.

To conclude, seminal studies have revealed that large, Poisson-like variability can spontaneously emerge from the collective dynamics of completely deterministic model neurons that have been arranged in random architectures. As the mammalian cortex consists of large and intricate assemblies of neurons, it appears reasonable to hypothesize that collective network effects might significantly contribute to the total neural variability that has been measured from data.

1.1.2 Point-process and firing rate variability

If neurons in balanced cortical networks behave as Poisson spike generators, one could rationally hypotesize that cortical cells mostly encode information through the firing rate variables which control their irregular spiking. While the network receives and processes its inputs, firing rates could be stationary or varying in time. As in Fig. 1.1 **a**, it is tempting to try to estimate the time course of a single cell firing rate by averaging the neural activity across many repetitions of the same measurement. Such approach, that has been widely adopted in the literature, in fact only returns a measure of the trial-averaged firing rate.

More recently, the necessity of considering inter-trial rate fluctuations as well has been pointed out [140, 35, 31]. Isolating the variability which derives from variations in the firing rates, in fact, can help build a more precise mapping between neural activity and behaviour, especially in tasks where the behavioural output is variable itself [34, 33, 49, 102]. In some cases, furthermore, the cortical response to behavioural stimuli might be not evident at the level of the average firing rate, while it might appear at the level of the amplitude of single-trial fluctuations [35]. Finally, it has been suggested that a principled analysis of rate variability could help distinguishing between alternative computational models for cortical dynamics [35, 31].

When the neural activity is averaged across trials, the variability in the firing rates, or *gain* variability, is washed out together with the variability associated with the Poisson spike generation, also referred to as *point-process* variability. Both sources of irregularity, on the other hand, are integrated together when standard variability measures are applied, like the Fano factor for the spike count and the coefficient of variation for the inter-spike intervals.

The two contributions can be disentangled by considering doubly stochastic models of spike initiation [35, 31, 56]. In those frameworks, spikes are emitted at random from an underlying time-varying firing rate. Crucially, the rate consists of a deterministic, stimulus-driven component ν, which is frozen across different trials, and of a trial-dependent gain G which enters multiplicatively [56]:

$$\phi = G\nu. \tag{1.3}$$

Introducing the gain variable G allows to largely improve the precision of the fit to the neural data recorded from several distinct cortical regions [56]. As shown in Fig. 1.3 **a**, furthermore, assuming a multiplicative dependence of the rate on the gain predicts an excess of variance which resembles the super-Poisson variability which has been long observed in the literature [126, 120, 140]. The time traces of the gain that are directly inferred from the neural data display indeed large trial-to-trial variability, which can be measured in terms of its coefficient of variation (Fig. 1.3 **b**). Larger firing rate variability is typically observed in areas which are higher in the cortical hierarchy.

The gain auto-correlogram in Fig. 1.3 **c** reveals that rate fluctuations are typically slow, with relaxation time scales which can last up to several minutes. On average, the cortical relaxation time scales follow a precise hierarchical ordering, with sensory and prefrontal areas exhibiting, respectively, shorter and longer time constants [92].

A significative fraction of the firing rate variability appears to be shared across neurons covering wide cortical areas, and is likely to importantly contribute to the correlation patterns between pairs of cells [56, 35]. This observation is agreement with the long-standing hypothesis that rate fluctuations derive from modulations in the excitability controlled by top-down signals linked to arousal and attention [140].

On the other hand, the shared component of rate variability seems to be restricted to relatively fast fluctuations [56]. The tails of the auto-correlation function which correspond to slow fluctuations, indeed, are absent in the cross-correlogram measured within pairs of different neurons (Fig. 1.3 **c**). Single neurons thus seem to develop local and extremely slow firing rate modulations, a scenario which is more difficult to reconcile with a top-down modulatory hypothesis. Fast and slow firing rate modulations could thus originate from distinct generating mechanisms.

1.2 Chaotic regimes in networks of firing rate units

One recent hypothesis suggests that, similarly to point-process variability, slow and local firing rate modulations could be produced intrinsically through the recurrent circuitry of the cortex [66]. A convenient network model, which spontaneously sustains slowly fluctuating dynamical regimes, was found almost thirty years ago in a seminal work by Sompolinsky and colleagues [127].

In this study, the collective behaviour of large disordered networks of non-linear units is examined (Fig. 1.4 **a**). Every node in the network is characterized by a continuous state variable, whose dynamics obeys a smooth evolution law. At every node, the activation variable

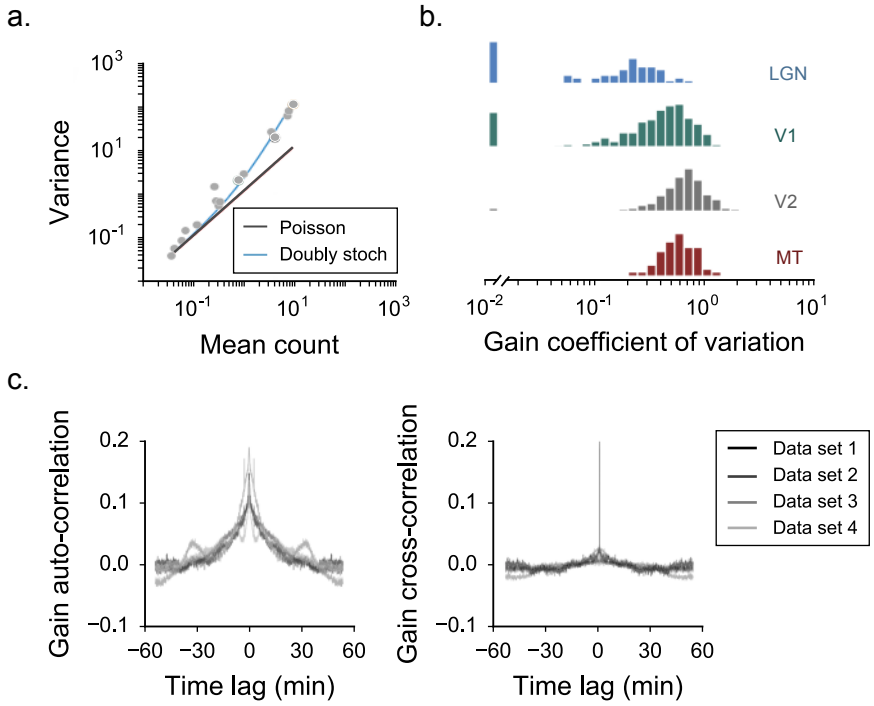


FIGURE 1.3: Spike trains from cortical in-vivo recordings can be parsimoniously described by a doubly stochastic model, where the firing rate across different trials contains both a frozen and a variable component (Eq. 1.3). **a.** Spike count variance-to-mean relation for a single V1 neuron stimulated with grating stimuli drifting in different directions. The prediction from the doubly stochastic model fitted to data is displayed in blue. The prediction from a simple Poisson model is shown in black. Means and variances were computed over 125 repetitions of each stimulus direction. **b.** Trial-to-trial variability of the gain variable inferred from the doubly stochastic model, quantified through its coefficient of variation. The full distribution is obtained by performing the analysis on different time windows. Different colors refer to different data sets, recorded from different cortical areas in the visual pathway. **c.** Temporal structure of the inferred gain, measured in four different data sets and averaged across many recorded cells. Left: auto-correlation, right: cross-correlation. Adapted from [56].

x_i, loosely interpreted as the total current entering the unit, is non-linearly transformed into an output $\phi(x_i)$. Critically, the network architecture is random, i.e. the pairwise coupling parameters are drawn from a Gaussian probability distribution.

It was found that the overall strength of the network connectivity structure controls the appearance of smooth chaotic fluctuations from a silent network regime. At the bifurcation point, an extensive number of eigenvectors become unstable. Network activity is then push-pulled in many different random directions, resulting in irregular and uncorrelated fluctuations with complex spatio-temporal profiles (Fig. 1.4 **b**). In this irregular state, an explicit calculation of the Lyapunov exponents indicates that fluctuations have exponentially short memory of their past history. As the dynamics is fully deterministic, network activity is formally chaotic.

A second attractive feature of this network model is the rich variety of activity time scales that it can support. The time scale of chaotic fluctuations is indeed controlled by the main

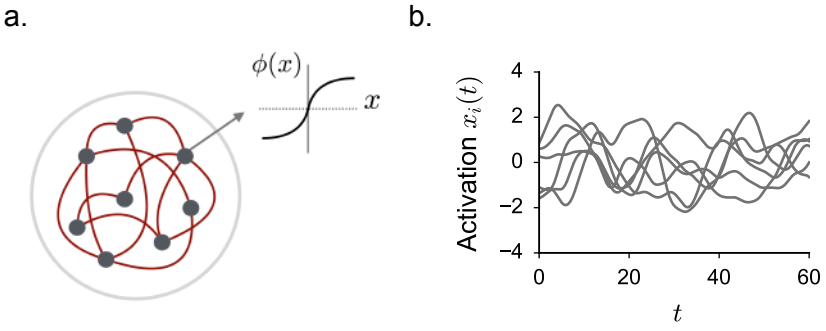


Figure 1.4: Irregular and smooth fluctuations spontaneously emerge as collective phenomenon in networks of randomly coupled non-linear units. **a.** A random network: activity in each node is described by a continuous variable $x_i(t)$ which obeys a smooth temporal evolution law. The non-linear input-output transformation performed by single units is modeled through the activation function $\phi(x)$. **b.** Sample of chaotic activity in finite networks: temporal evolution of the activation variable x_i for six randomly selected unit. Activity traces fluctuate irregularly around zero.

parameter of the system, i.e. the overall connectivity strength. Close to the critical point, the time decay of fluctuations sharply increases; it formally diverges in the limit of infinite networks size.

Although the network dynamics in [127] can be formally mapped into a standard firing rate model [42, 151], whether similar irregular states are expected to appear in biologically-motivated models of cortical networks has been a long-standing question. Indeed, while this model captures the essence of the non-linear input-output transformation taking place in real neurons, it lacks the elementary features that would enable a direct comparison with other more realistic networks models. For many years, very little effort has been devoted to build and explore such a link.

In the last few years, this class of models has been attracting increasing attention. As already discussed, this renewed interest can be partially attributed to the recent improvement in the neural recordings techniques, that have allowed to systematically disentangle multiple levels of variability in the neural data.

A second significant contribution has come from the recent advances in the research field at the frontier between neuroscience and machine learning. The model from [127] is indeed widely adopted in novel lines of research which explore learning and plasticity mechanisms in recurrent network models [67, 132, 73, 85, 138]. Because of the complex temporal dependencies that are intrinsically generated by recurrent circuits, training recurrent neural networks has been historically a cumbersome task. A variety of novel training strategies have only recently allowed to efficiently build artificial computational models which can solve elaborate spatio-temporal tasks. Apart from being attractive tools for the theoretical and machine learning communities, trained networks have been combined with neural recordings to get valuable insights on the dynamics and coding principles of the cortex [83, 15, 100, 128, 88].

Despite their efficacy and prediction power, trained neural networks are in most of the cases black box models designed on obscure – and hard to capture – dynamical principles [13, 133]. Most of the modelling efforts consecrated to understanding large circuits dynamics

have been focusing indeed on completely random network architectures [24, 145, 127], where the relationship between connectivity and dynamics can be understood in great detail. On the contrary, a rigorous understanding of non-random computational networks presents, from a theoretical point of view, several novel challenges which still need to be addressed. As a result, what are the dynamical mechanisms underlying computations in trained networks – and how variability is characterized in computational circuit models – are still largely unsolved theoretical questions.

1.3 Outline of the work

In this thesis, we investigate how intrinsically generated variability can be integrated in more realistic models of cortical networks which can serve as elementary units of computation. The dissertation collects the results of three years of work and consists of three main parts.

I. In the first part, we take direct inspiration from the original work in [127] and we design a random network model which includes several additional constraints directly motivated by biology. We consider a network architecture which respects Dale's law: every unit in the network can either excite or inhibite his neighbours. Moreover, we include more realistic, positively defined, current-to-rate activation functions. We show that rate fluctuations appear in strongly coupled excitatory-inhibitory architectures, and we study how variability depends on further biophysical restrictions, like firing-rate saturation, heterogeneity in the connectivity and spiking noise. A constrained network model allows for a neater comparison with the more realistic networks of spiking units that have been traditionally adopted as models for irregular activity in cortical circuits. In this perspective, we show that our simple rate description can be used to help understanding the more complex dynamics generated in strongly inter-connected networks of leaky integrate-and-fire neurons, where classical mean field approaches fail to provide a good description of spiking activity [95].

II. In the second part, we turn to a simple computational architecture, which includes a random network together with a single feedback signal. Such elementary connectivity scheme has been successfully exploited in different training setups [67, 132]. Since the network connectivity is not far from being completely random, we show that the classical mathematical tools developed for the analysis of large disordered systems can be fruitfully applied to this scenario. We perform a systematic analysis of a network model designed to behave as a generator of arbitrary periodic patterns. Consistently with our theory, we show that training performance significantly improves in highly disordered architectures. When the random component of the connectivity is strong, in fact, the intrinsic heterogeneity prevents network activity from synchronizing. Learning can thus take advantage of a widely decorrelated set of neural activitiy from which the desired output pattern can be solidly reconstructed.

III. Random networks with a single feedback unit can be more generally thought as novel recurrent architectures where the global connectivity structure consists of the sum of a random and of a unit rank structured term. More in general, the idea that neural computations in large recurrent networks only rely on low-dimensional connectivity structures is widely shared across several training frameworks [65, 67, 132, 20, 48]. Motivated by this observation, in the third part of this dissertation we explore the dynamics of large random networks per-

turbed by weak, low-dimensional connectivity structures. We find that rank one and rank two connectivity structures that are generated by high-dimensional random vectors can generate a rich variety of irregular and heterogeneous dynamical regimes. The knowledge of the stable states of the dynamics allows to easily design partially structured models which can perform simple tasks. In the resulting computational models, our theoretical framework allows to predict the variability properties and the largest relaxation time scales of the dynamics. It further permits to neatly interpret the low-dimensional evolution of the population activity in terms of the few structured directions that are specified by the network architecture and inputs.

Part I

INTRINSICALLY-GENERATED FLUCTUATIONS IN RANDOM NETWORKS OF EXCITATORY-INHIBITORY UNITS

Summary of Chapters 2 - 3 - 4

Recurrent networks of non-linear units display a variety of dynamical regimes depending on the structure of their synaptic connectivity. A particularly remarkable phenomenon is the appearance of strongly fluctuating, chaotic activity in networks of deterministic, but randomly connected rate units. How this type of intrinsically generated fluctuations appears in more realistic networks of spiking neurons has been a long standing question. The comparison between rate and spiking networks has in particular been hampered by the fact that most previous studies on randomly connected rate networks focused on highly simplified models, in which excitation and inhibition were not segregated and firing rates fluctuated symmetrically around zero because of built-in symmetries.

To ease the comparison between rate and spiking networks, we investigate the dynamical regimes of sparse, randomly-connected rate networks with segregated excitatory and inhibitory populations, and firing rates constrained to be positive.

Extending the dynamical mean field theory, we show that network dynamics can be effectively described through two coupled equations for the mean activity and the auto-correlation function. As a consequence, we identify a new signature of intrinsically generated fluctuations on the level of mean firing rates. We moreover found that excitatory-inhibitory networks develop two different fluctuating regimes: for moderate synaptic coupling, recurrent inhibition is sufficient to stabilize fluctuations; for strong coupling, firing rates are stabilized solely by the upper bound imposed on activity. These results extend to more general network architectures, and to rate networks receiving noisy inputs mimicking spiking activity. Finally, we show that signatures of those dynamical regimes appear in networks of integrate-and-fire neurons.

A substantial fraction of this part of the dissertation is adapted from the manuscript: *Intrinsically-generated fluctuating activity in excitatory-inhibitory networks* by F. Mastrogiuseppe and S. Ostojic, PLoS Computational Biology (2017) [87].

Dynamical Mean Field description of excitatory-inhibitory networks

2

In the first part of this dissertation, we study how the transition to a chaotic, slowly fluctuating dynamical regime, which was first observed in [127], translates to more realistic network models. We design a non-linear firing rate network which includes novel mathematical contraints motivated by biology, and we quantitatively address its spontaneous dynamics.

If the synaptic coupling is globally weak, firing rate networks can be described with the help of standard approaches from dynamical systems theory, like linear stability analysis. However, in the strong coupling regime, a rigorous description of self-sustained fluctuations can be derived only at the statistical level. To this end, we adopt and extend the theoretical framework first proposed in [127], which provides an adequate description of irregular temporal states. In such approach, irregular trajectories are thought as random processes sampled from a continuous probability distribution, whose first two momenta can be computed self-consistently [127]. This technique, commonly referred to as Dynamical Mean Field (DMF) theory, has been inherited from the study of disordered systems of interacting spins [40], and provides a powerful and flexible tool for understanding dynamics in disordered rate networks.

In this chapter, we adapt this approach to the study of more realistic excitatory-inhibitory network models. We derive the mean field equations which will become the central core of the analysis which is carried out in details in the rest of part I. To begin with, we review the methodology of DMF, and we present the results that such theory implies for the original, highly symmetrical model. This first section is effectively a re-interpretation of the short paper by [127]. In the second section, we introduce and motivate the more biologically-inspired model that we aim at studying, and we show that an analogous instability from a fixed point to chaos can be predicted by means of linear stability arguments. In order to provide an adequate self-consistent description on the irregular regime above the instability, we extend the DMF framework to include non-trivial effects due to non-vanishing first-order statistics.

2.1 Transition to chaos in recurrent random networks

The classical network model in [127] is defined through a non-linear continuous-time dynamics which makes it formally equivalent to a traditional firing rate model [151, 42]. Firing rate models are meant to provide a high-level description of circuit dynamics, as spiking activity is averaged over one or more degrees of freedom to derive a simpler description in terms of smooth state variables. From a classical perspective, firing rate units provide a good description of the average spiking activity in small neural populations. Equivalently, they can well approximate the firing of single neurons if the synaptic filtering time-scale is large enough. Although in this chapter we don't focus on any specific interpretation, a sloppy terminology where the words *unit* and *neuron* are used indifferently will be adopted.

The state of each unit in the network is described through an activation variable x_i which is commonly interpreted as the net input current entering the cell. The current-to-rate transformation that is performed in spiking neurons is modeled through a monotonically increasing function ϕ, such that the variable $\phi(x_i)$ represents the instantaneous output firing rate of the unit.

As the network consists of many units ($i = 1, ..., N$), the current entering neuron i includes many contributions, whose values are proportional to the firing rate of the pre-synaptic neurons. The strength of the synapse from neuron j to neuron i is modeled through the connectivity parameter J_{ij}. The coupled dynamics obey the following temporal evolution law:

$$\dot{x_i}(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij}\phi(x_j(t)). \tag{2.1}$$

The first contribution on the r.h.s. is a leak term, which ensures the activation variables x_i decays back to baseline in absence of any forcing current. The incoming contributions from other units in the network sum linearly. Note that we have rescaled time to set the time constant to unity.

In the paper by [127], the authors focus on a random all-to-all Gaussian connectivity (Fig. 2.1 **a**). We thus have $J_{ij} = g\chi_{ij}$ with $\chi_{ij} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1/N)$. Such scaling for the variance ensures that single units experience finite input fluctuations even in the limit of very large networks. The parameter g controls the global strength of synaptic coupling. As neurons can make both excitatory and inhibitory connections, this connectivity scheme does not respect Dale's law. The activation function is a symmetric sigmoid ($\phi(x) = \tanh(x)$), which takes positive and negative values. In the original network, furthermore, neither constant nor noisy external inputs are considered.

As we will show in the next sections, all those elements together result in an extremely simplified dynamics, where the transition to chaos can be only measured at the level of the second-order statistics of the network activity distribution.

2.1.1 Linear stability analysis

To begin with, we notice that the model admits an homogeneous stationary solution for which the network is completely silent: $x_i^0 = 0\ \forall i$. For a fixed, randomly chosen connectivity matrix, the network we consider is fully deterministic, and can therefore be examined using standard dynamical system techniques [131]. We thus derive a first intuitive picture about the network dynamics by evaluating the linear stability of the homogeneous fixed point.

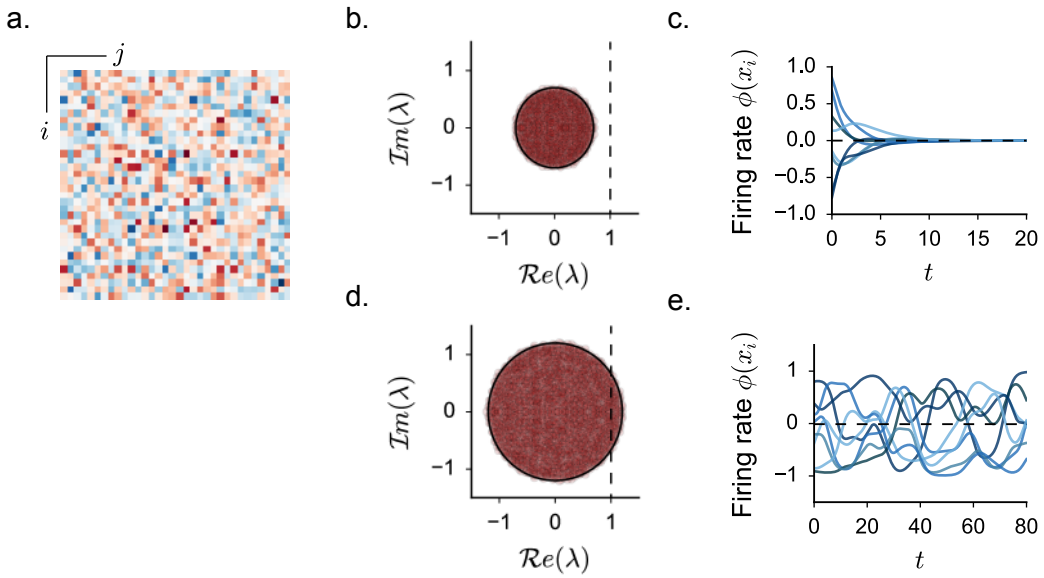


FIGURE 2.1: Linear stability analysis and transition to chaos in all-to-all Gaussian networks with symmetric activation function [127]. **a.** The Gaussian connectivity matrix $g\chi_{ij}$. **b-c.** Stationary regime: $g = 0.8$. In **b**: eigenspectrum of the stability matrix S_{ij} for a simulated network of $N = 2000$ units. In good agreement with the circular law prediction, the eigenvalues lie in a compact circle of radius g (continuous black line). Dashed line: instability boundary. In **c**: sample of simulated activity for eight randomly chosen units. **d-e.** Chaotic regime: $g = 1.2$. Same figures as in **b-c**.

The linear response of the system when pushed away from the fixed point can be studied by tracking the time evolution of a solution in the form: $x_i(t) = x_0 + \delta x_i(t)$. Close to the fixed point, the function $\phi(x)$ can be expanded up to the linear order in $\delta x_i(t)$. This results in a system of N coupled linear differential equations, whose dynamical matrix is given by: $S_{ij} = \phi'(0)g\chi_{ij} - \delta_{ij}$. Note that here $\phi'(0) = 1$.

As a result, the perturbation $\delta x_i(t)$ will be re-absorbed if $\mathcal{R}e(\lambda_i) < 1$ for all i, λ_i being the ith eigenvalue of the asymmetric random matrix $g\chi_{ij}$. We are thus left with the problem of evaluating the eigenspectrum of a Gaussian random matrix. If one focuses on very large networks, the circular (or Girko's) law can be applied [54, 136]: the eigenvalues of $g\chi_{ij}$ lie in the complex plane within a circular compact set of radius g. Although its prediction is exact only in the thermodynamic limit ($N \to \infty$), the circular law also reasonably well approximate the eigenspectrum of finite random matrices.

We derive that, at low coupling strength ($g < 1$), the silent fixed point is stable (Fig. 2.1 **b-c**). More than this, $x_0 = 0$ is a global attractor, as S_{ij} is a contraction [146]. Numerical simulations confirm that, in this parameter region, network activity settles into the homogeneous fixed point. For $g > 1$, the fixed point is unstable, and the network exhibits ongoing dynamics in which single neuron activity fluctuates irregularly both in time and across different units (Fig. 2.1 **d-e**). As the system is deterministic, these fluctuations are generated intrinsically by strong feedback along unstable modes, which possess a random structure inherited from the random connectivity matrix.

2.1.2 The Dynamical Mean Field theory

The non-stationary regime cannot be easily analyzed with the tools of classical dynamical systems. To this end, the authors in [127] adopted a mean field approach to develop an effective statistical description of network activity. In this section, we propose a review of such technique; our analysis is based on [127] and subsequent works [99, 141, 89].

Rather than attempting to describe single trajectories, the main idea is to focus on their statistics, which can be determined by averaging over different initial conditions, time and the different instances of the connectivity matrix. Dynamical Mean Field (DMF) acts by replacing the fully deterministic interacting network by an equivalent stochastic system. More specifically, as the interaction between units $\sum_j J_{ij}\phi(x_j)$ consists of a sum of a large number of terms, it can be replaced by a Gaussian stochastic process $\eta_i(t)$. Such a replacement provides an exact mathematical description under specific assumptions on the chaotic nature of the dynamics [16, 90] in the limits of large network size N. In this thesis, we will treat it as an approximation, and we will assess the accuracy of this approximation by comparing the results with simulations performed for fixed N.

Replacing the interaction terms by Gaussian processes transforms the system into N identical Langevin-like equations:

$$\dot{x_i}(t) = -x_i(t) + \eta_i(t). \tag{2.2}$$

As $\eta_i(t)$ is a Gaussian noise, each trajectory $x_i(t)$ emerges thus as a Gaussian stochastic process. As we will see, the stochastic processes corresponding to different units become uncorrelated and statistically equivalent in the limit of a large network, so that the network is effectively described by a single process.

Within DMF, the mean and correlations of this stochastic process are determined self-consistently, by requiring that averages over $\eta_i(t)$ be identical to averages over time, instances of the connectivity matrix and initial conditions in the original system. Both averages will be indicated with []. For the mean, we get:

$$[\eta_i(t)] = g[\sum_{j=1}^{N} \chi_{ij}\phi(x_j(t))] = g\sum_{j=1}^{N}[\chi_{ij}][\phi(x_j(t))] = 0 \tag{2.3}$$

as $[\chi_{ij}] = 0$. In the second equality, we assumed that activity of different units decorrelates in large networks; in particular, that activity of unit j is independent of its outgoing connections J_{ij}. As we will show in few lines, this assumption is self-consistent. In the mathematical literature, it has been referred to as *local chaos* hypothesis [8, 52, 90].

The second-order statistics of the effective input gives instead:

$$\begin{aligned}[\eta_i(t)\eta_j(t+\tau)] &= g^2[\sum_{k=1}^{N}\chi_{ik}\phi(x_k(t))\sum_{l=1}^{N}\chi_{jl}\phi(x_l(t+\tau))] \\ &= g^2\sum_{k=1}^{N}\sum_{l=1}^{N}[\chi_{ik}\chi_{jl}][\phi(x_k(t))\phi(x_l(t+\tau))].\end{aligned} \tag{2.4}$$

As $[\chi_{ik}\chi_{jl}] = \delta_{ij}\delta_{kl}/N$, cross-correlations vanish, while the auto-correlation results in:

$$[\eta_i(t)\eta_i(t+\tau)] = g^2[\phi(x_i(t))\phi(x_i(t+\tau))]. \tag{2.5}$$

We will refer to the firing rate auto-correlation function $[\phi(x_i(t))\phi(x_i(t+\tau))]$ as $C(\tau)$. Consistently with our starting hypothesis, the first- and the second-order statistics of the Gaussian process are uncorrelated from one unit to the other.

Once the probability distribution of the effective input has been characterized, we derive a statistical description of the network activity in terms of the activation variable $x_i(t)$ by solving the Langevin equation in Eq. 2.2.

Trivially, the first-order statistics of $x_i(t)$ and $\eta_i(t)$ asymptotically coincide, so that the mean input always vanishes. In order to derive the auto-correlation function $\Delta(\tau) = [x_i(t)x_i(t+\tau)]$, we derive twice with respect to τ and we combine Eqs. 2.2 and 2.5 to get the following time evolution law:

$$\ddot{\Delta}(\tau) = \Delta(\tau) - g^2 C(\tau). \tag{2.6}$$

We are thus left with the problem of writing down an explicit expression for the firing rate auto-correlation function $C(\tau)$. To this aim, we write $x(t)$ and $x(t+\tau)$ as Gaussian variables which obey $[x(t)x(t+\tau)] = \Delta(\tau)$ and $[x(t)^2] = [x(t+\tau)^2] = \Delta_0$, where we defined the input variance $\Delta_0 = \Delta(\tau = 0)$. One possible choice is:

$$\begin{aligned} x(t) &= \sqrt{\Delta_0 - |\Delta(\tau)|}x_1 + \sqrt{|\Delta(\tau)|}z \\ x(t+\tau) &= \sqrt{\Delta_0 - |\Delta(\tau)|}x_2 + \mathrm{sgn}(\Delta(\tau))\sqrt{|\Delta(\tau)|}z \end{aligned} \tag{2.7}$$

where x_1, x_2 and z are Gaussian variables with zero mean and unit variance. For reasons which will become clear in few steps, we focus on the case $\Delta(\tau) > 0$. Under this assumption, the firing rate auto-correlation function can be written as:

$$C(\tau) = \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\sqrt{\Delta_0 - \Delta(\tau)}x + \sqrt{\Delta(\tau)}z) \right]^2 \tag{2.8}$$

where used the short-hand notation: $\int \mathcal{D}z = \int_{-\infty}^{+\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz$.

From a technical point of view, Eq. 7.61 is now a second-order differential equation, whose time evolution depends on its initial condition Δ_0. This equation admits different classes of solutions which are in general hard to isolate in an explicit form. Luckily, we can reshape our problem into a simpler, more convenient formulation.

Isolating the solutions We observe that Eq. 7.61 can be seen as analogous to the equation of motion of a classical particle in a one-dimensional potential:

$$\ddot{\Delta} = -\frac{\partial V(\Delta, \Delta_0)}{\partial \Delta} \tag{2.9}$$

The potential $V(\Delta, \Delta_0)$ is given by an integration over Δ:

$$V(\Delta, \Delta_0) = -\int_0^{\Delta} d\Delta' \left[\Delta' - g^2 C(\Delta', \Delta_0) \right]. \tag{2.10}$$

One can check that this results in:

$$V(\Delta, \Delta_0) = -\frac{\Delta^2}{2} + g^2 \int \mathcal{D}z \left[\int \mathcal{D}x \Phi(\sqrt{\Delta_0 - \Delta}x + \sqrt{\Delta}z) \right]^2 \tag{2.11}$$

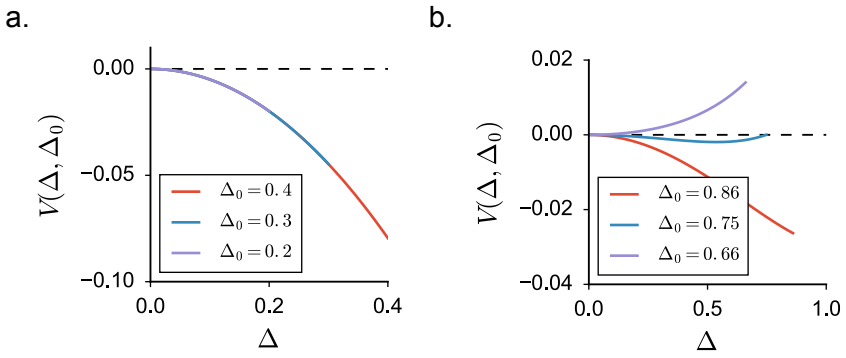


Figure 2.2: Shape of the potential $V(\Delta, \Delta_0)$ for different initial conditions Δ_0. **a.** Weak copling regime: $g < g_C$. **b.** Strong coupling regime: $g > g_C$.

where $\Phi(x) = \int_{-\infty}^{x} \phi(x')\,\mathrm{d}\,x'$. In the present framework, $\Phi(x) = \ln(\cosh(x))$.

In absence of external noise, the initial condition to be satisfied is $\dot{\Delta}(\tau = 0) = 0$, which implies null kinetic energy for $\tau = 0$. A second condition is given by: $\Delta_0 > |\Delta(\tau)|\ \forall \tau$. The solution $\Delta(\tau)$ depends on the initial value Δ_0, and is governed by the energy conservation law:

$$V(\Delta(\tau = 0), \Delta_0) = V(\Delta(\tau = \infty), \Delta_0) + \frac{1}{2}\dot{\Delta}(\tau = \infty)^2 \tag{2.12}$$

The stationary points and the qualitative features of the $\Delta(\tau)$ trajectory depend then on the shape of the potential V. We notice that for the symmetric model from [127], the derivative of the potential in $\Delta = 0$ always vanishes, suggesting a possible equilibrium point. The full shape of V is determined by the values of g and Δ_0. In particular, a critical value g_C exists such that:

- when $g < g_C$, the potential has the shape of a concave parabola centered in $\Delta = 0$ (Fig. 2.2 **a**). The only bounded solution is $\Delta = \Delta_0 = 0$;
- when $g > g_C$, the potential admits different qualitative configurations and an infinite number of different $\Delta(\tau)$ trajectories. In general, the motion in the potential will be oscillatory (Fig. 2.2 **b**).

We conclude that, in the weak coupling regime, the only acceptable solution is centered in 0 and has vanishing variance. In other terms, in agreement with our linear stability analysis, we must have $x_i(t) = 0\ \forall t$.

In the strong coupling regime, we observe that a particular solution exists, for which $\Delta(\tau)$ decays to 0 as $\tau \to \infty$. In this final state, there is no kinetic energy left. For this particular class of solutions, Eq. (2.12) reads:

$$V(\Delta_0, \Delta_0) = V(0, \Delta_0). \tag{2.13}$$

More explicitly, we have:

$$\frac{\Delta_0^2}{2} = g^2 \left\{ \int \mathcal{D}z \Phi^2(\sqrt{\Delta_0} z) - \left(\int \mathcal{D}z \Phi(\sqrt{\Delta_0} z) \right)^2 \right\}. \tag{2.14}$$

In the following, we will often use the compact notation:

$$\frac{\Delta_0^2}{2} = g^2 \left\{ [\Phi^2] - [\Phi]^2 \right\}. \tag{2.15}$$

A monotonically decaying auto-correlation function implies a dynamics which loses memory of its previous states, and is compatible with a chaotic state. In the original study by Sompolinsky et al. [127], the average Lyapunov exponent is computed. It is shown that the monotonically decreasing solution is the only self-consistent one, as the correspondent Lyapunov exponent is positive.

Once Δ_0 is computed through Eq. 2.15, its value can be injected into Eq. 7.61 to get the time course of the auto-correlation function. The decay time of $\Delta(\tau)$, which depends on g, gives an estimation of time scale of chaotic fluctuations. As the transition in g_C is smooth, the DMF equations can be expanded around the critical coupling to show that such time scale diverges when approaching the transition from above. Very close to $g = g_C$, the network can thus support arbitrarily slow spontaneous activity.

Numerical inspection of the mean field solutions suggest that, as it has been predicted by the linear stability analysis, $g_C = 1$ (Fig. 2.3 **b**). This can be also rigorously checked by imposing that, at the transition point, the first and the second derivative of the potential vanish in $\Delta = 0$.

To conclude we found that, above $g = 1$, the DMF predicts the emergence of chaotic trajectories which fluctuate symmetrically around 0. In the large network limit, different trajectories behave as totally uncoupled processes. Their average amplitude can be computed numerically as solution of the non-linear self-consistent equation in 2.15.

2.2 Fast dynamics: discrete time evolution

As a side note, we briefly consider a closely related class of models which has been extensively adopted in the DMF literature. In this formulation, the dynamics is given by a discrete time update:

$$x_i(t+1) = \sum_{j=1}^{N} J_{ij}\phi(x_j(t)) \tag{2.16}$$

As there are no leak terms, fluctuations in the input current occur on an extremely fast timescale (formally, within one time step). All the other elements of the model, including J_{ij} and $\phi(x)$, are taken as in [127].

This discrete-time formulation has been used, for instance, in the first attempts to exploit random network dynamics for machine learning purposes [67]. It has also been adopted in several theoretically oriented studies, as analysing fast dynamics has two main advantages: mean field descriptions are easier to derive [89, 141, 30], and, in finite-size networks, the quasi-periodical route to chaos can be directly observed [46, 30, 4].

While finite size analysis falls outside the scope of this dissertation, we briefly review how the mean field equations adapt to discrete-time networks and how this description fits in the more general DMF framework.

Similar to the continuous-time case, the discrete-time dynamics admits an homogeneous fixed point in $x_0 = 0$. Furthermore, as it can be easily verified, the linear stability matrix of this stationary state coincides with $S_{ij} = g\chi_{ij} - \delta_{ij}$, so that an instability occurs in $g = 1$. In

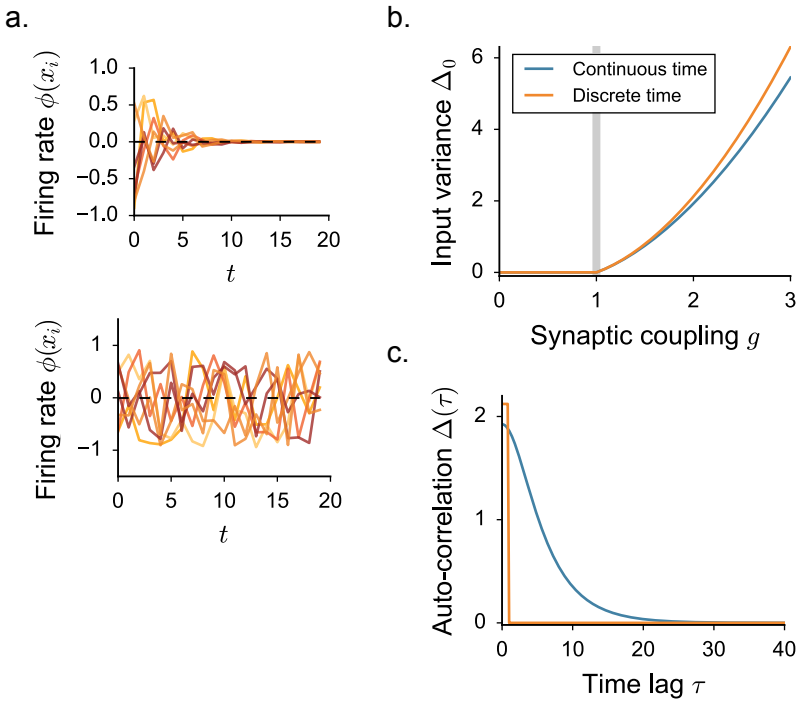


Figure 2.3: Discrete-time dynamics in random neural networks. **a.** Sample of simulated activity: time traces for eight randomly chosen units in the static (top) and in the chaotic regime (bottom). **b.** Bifurcation diagram for the second-order statistics Δ_0 as a function of the coupling strength parameter g. The value of Δ_0 is evaluated by solving iteratively the DMF equations for continuous- (Eq. 2.15) and discrete-time (Eq. 2.20) networks. Vertical line: critical coupling in $g_C = 1$. **c.** Temporal shape of the auto-correlation function for fixed $g = 1.3$. The time scale of discrete-time dynamics is set arbitrarily.

order to analyze dynamics beyond the instability, we apply DMF arguments. When defining the effective input $\eta_i(t) = \sum_{j=1}^{N} J_{ij}\phi(x_j(t))$, fast dynamics will translate in the following simple update rule:

$$x_i(t+1) = \eta_i(t) \tag{2.17}$$

where, at each time step, x_i is simply replaced by the stochastic effective input. As a consequence, by squaring and averaging over all the sources of disorder, we find that the input current variance obeys the following time evolution:

$$\Delta_0(t+1) = [\eta_i^2(t)] = g^2[\phi^2(t)]. \tag{2.18}$$

In the last equality, we used the self-consistent expression for the second-order statistics of η_i, which can be computed as in the continuous-time case, yielding to the same result. By expressing $x(t)$ as Gaussian variable, the evolution law for Δ_0 can be made explicit:

$$\Delta_0(t+1) = g^2 \int \mathcal{D}z\phi^2(\sqrt{\Delta_0(t)}z). \tag{2.19}$$

At equilibrium, the value of Δ_0 satisfies the fixed-point condition:

$$\Delta_0 = g^2 \int \mathcal{D}z\phi^2(\sqrt{\Delta_0(t)}z). \tag{2.20}$$

As it can be easily checked, this equation is satisfied for $\Delta_0 = 0$ when $g < 1$, while it admits a nontrivial positive solution above $g_C = 1$, corresponding to a fast chaotic phase (Fig. 2.3 **b**).

The solution that we derive from solving Eq. 2.20 does not coincide exactly with the solution we obtained in the case of continuous-time networks, although they share many qualitative features (Fig. 2.3 **b**). In contrast to discrete-time units, neurons with continuous-time dynamics act as low-pass filters of their inputs. For this reason, continuous-time chaotic fluctuations are characterized by a slower time scale (Fig. 2.3 **c**) and a smaller variance Δ_0 (Fig. 2.3 **b**).

We conclude this paragraph with a technical remark: our new equation for Δ_0 (Eq. 2.20) coincides with the general expression for stationary solutions in continuous-time networks. The latter can be derived from the continuous-time DMF equation $\ddot{\Delta}(\tau) = \Delta(\tau) - g^2 C(\tau)$ by setting $x(t) = x(t+\tau)$ and thus $\ddot{\Delta}(\tau) = 0$. From the analysis we just carried out, we conclude that the general stationary solution for continuous-time networks admits, together with the homogeneous fixed point, a non-homogeneous static branch for $g > 1$. As it is characterized by positive Lyapunov exponents, this solution is however never stable for continuous-time networks. This sets a formal equivalence between chaotic discrete-time and stationary continuous-time solutions which does not depend on the details of the network model. For this reason, it will return back several times within the body of this dissertation.

2.3 Transition to chaos in excitatory-inhibitory neural networks

As widely discussed in Chapter 1, network models which spontaneously sustain slow and local firing rate fluctuations are of great interest in the perspective of understanding the large, super-Poisson variability observed from in-vivo recordings [120, 56].

Furthermore, the random network model in [127] has been adopted in many training frameworks as a proxy for the unspecialized substrate on which plasticity algorithms can be applied. The original computational architecture from Jaeger [67], know as *echo-state machine*, adopts the variant of the model characterized by discrete-time dynamics [89, 30]. In later years, several training procedures have been designed for continuous-time models as well [132, 73, 28].

A natural question we would like to address is whether actual cortical networks exhibit dynamical regimes which are analogous to rate chaos.

The classical network model analyzed in [127] and subsequent studies [132, 73, 99, 6, 7, 130] rely on several simplifying features that prevent a direct comparison with more biologically constrained models such as networks of spiking neurons. In particular, a major simplification is a high degree of symmetry in both input currents and firing rates. Indeed, in the classical model the synaptic strengths are symmetrically distributed around zero, and excitatory and inhibitory neurons are not segregated into different populations, thus violating Dale's law. The current-to-rate activation function is furthermore symmetric around zero, so that the dynamics are symmetric under sign reversal. As a consequence, the mean activity in the network is always zero, and the transition to the fluctuating regime is characterized solely in terms of second order statistics.

To help bridge the gap between the classical model and more realistic spiking networks [24, 95], recent works have investigated fluctuating activity in rate networks that include additional biological constraints [95, 69, 58], such as segregated excitatory-inhibitory populations, positive firing rates and spiking noise [69]. In general excitatory-inhibitory networks,

the DMF equations can be formulated, but are difficult to solve, so that these works focused mostly on the case of purely inhibitory networks. These works therefore left unexplained some phenomena observed in simulations of excitatory-inhibitory spiking and rate networks [95], in particular the observation that the onset of fluctuating activity is accompanied by an elevation of mean firing rate.

Here we investigate the effects of excitation on fluctuating activity in inhibition-dominated excitatory-inhibitory networks [142, 91, 3, 111, 60, 61]. To this end, we focus on a simplified network architecture in which excitatory and inhibitory neurons receive statistically identical inputs [24]. For that architecture, dynamical mean field equations can be fully solved.

2.3.1 The model

We consider a large, randomly connected network of excitatory and inhibitory rate units. Similarly to [127], the network dynamics are given by:

$$\dot{x_i}(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij}\phi(x_j(t)) + I_i. \tag{2.21}$$

In some of the results which follow, we will include a fixed or noisy external current I_i. The function $\phi(x)$ is a monotonic, positively defined activation function that transforms input currents into output activity.

For the sake of simplicity, in most of the applications we restrict ourself to the case of a threshold-linear activation function with an offset γ. For practical purposes, we take:

$$\phi(x) = \begin{cases} 0 & x < -\gamma \\ \gamma + x & -\gamma \leq x \leq \phi_{max} - \gamma \\ \phi_{max} & x > \phi_{max} - \gamma \end{cases} \tag{2.22}$$

where ϕ_{max} plays the role of the saturation value. In the following, we set $\gamma = 0.5$.

We focus on a sparse, two-population synaptic matrix identical to [24, 95]. We first study the simplest version in which all neurons receive the same number $C \ll N$ of incoming connections (respectively $C_E = fC$ and $C_I = (1-f)C$ excitatory and inhibitory inputs). More specifically, here we consider the limit of large N while C (and the synaptic strengths) are held fixed [9, 24]. We set $f = 0.8$.

All the excitatory synapses have strength J and all inhibitory synapses have strength $-gJ$, but the precise pattern of connections is assigned randomly (Fig. 2.4 **a**). For such connectivity, excitatory and inhibitory neurons are statistically equivalent as they receive statistically identical inputs. This situation greatly simplifies the mathematical analysis, and allows us to obtain results in a transparent manner. In a second step, we show that the obtained results extend to more general types of connectivity.

Our analysis largely builds on the methodology that we have been reviewing in the previous section for the case of the simpler network as in [127].

2.3.2 Linear stability analysis

As the inputs to all units are statistically identical, the network admits a homogeneous fixed point in which the activity is constant in time and identical for all units, given by:

$$x_0 = J(C_E - gC_I)\phi(x_0). \tag{2.23}$$

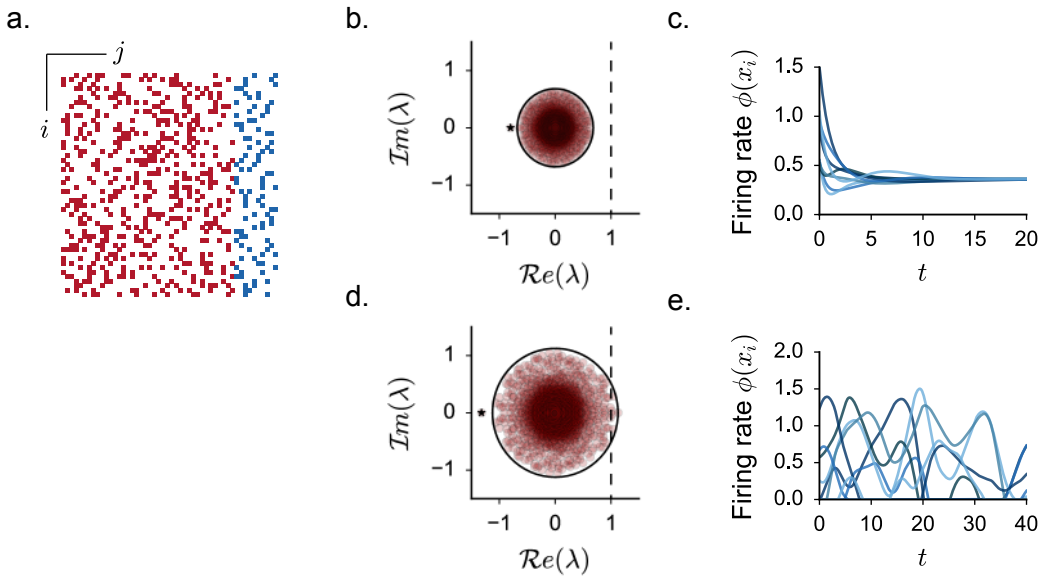


FIGURE 2.4: Linear stability analysis and transition to chaos in excitatory-inhibitory networks with threshold-linear activation function. **a.** The sparse excitatory-inhibitory connectivity matrix J_{ij}. **b-c.** Stationary regime: $J < J_0$. In **b**: eigenspectrum of the stability matrix S_{ij} for a simulated network of $N = 2000$ units. In good agreement with the circular law prediction, the eigenvalues lie in a compact circle of approximated radius $J\sqrt{C_E + g^2 C_I}$ (black continuous line). Black star: eigenspectrum outlier in $J(C_E - gC_I) < 0$. Dashed line: instability boundary. In **c**: sample of simulated activity for eight randomly chosen units. **d-e.** Chaotic regime: $J > J_0$. Same figures as in **b-c** .

The linear stability of this fixed point is determined by the eigenvalues of the matrix $S_{ij} = \phi'(x_0)J_{ij}$.

In the limit of large networks, the eigenspectrum of J_{ij} consists of a continuous part that is densely distributed in the complex plane over a circle of radius $J\sqrt{C_E + g^2 C_I}$, and of a real outlier given by the effective balance of excitation and inhibition in the connectivity $J(C_E - gC_I)$ (Fig. 2.4 **b-d**) [98, 54, 136, 135]. We focus here on an inhibition-dominated network corresponding to $g > C_E/C_I$. In this regime, the real outlier is always negative and the stability of the fixed point depends only on the continuous part of the eigenspectrum. The radius of the eigenspectrum disk, in particular, increases with the coupling J, and an instability occurs when the radius crosses unity. The critical coupling J_0 is given by:

$$\phi'(x_0)J_0\sqrt{C_E + g^2 C_I} = 1 \tag{2.24}$$

where x_0 depends implicitly on J through Eq. 2.23 and the gain $\phi'(x)$ is in general finite and non-negative for all the values of x.

Numerical simulations suggest that, above the instability, the positively-bounded firing rate trajectories undergo spatial and temporal irregular fluctuations. In order to provide a characterization of chaotic activity, we extend and adapt the DMF theory to the new architecture. As we will see, the main novelties derive from the necessity to include in the framework a self-consistent description of the non-vanishing first-order statistics.

2.3.3 Deriving DMF equations

DMF theory acts by replacing the fully deterministic coupling term $\sum_j J_{ij}\phi(x_j) + I$ in Eq. 2.21 by an equivalent Gaussian stochastic process η_i. Our aim is thus to compute self-consistently the first and second order moments of the effective noise η_i by averaging over time, units, initial conditions and realizations of the random matrix. For simplicity, we focus here on the case of constant and homogeneous inputs.

For the mean, we get:

$$\begin{aligned}[\eta_i(t)] &= \left[\sum_{j=1}^{N} J_{ij}\phi_j(t) + I\right] = \sum_{j_E=1}^{C_E} J[\phi_{j_E}] - g\sum_{j_I=1}^{C_I} J[\phi_{j_I}] + I \\ &= J(C_E - gC_I)[\phi] + I\end{aligned} \tag{2.25}$$

where the indices j_E and j_I run over the excitatory and the inhibitory units pre-synaptic to unit i. We moreover used the short-hand notation: $\phi_i(t) \coloneqq \phi(x_i(t))$. We assume that the mean values of x and ϕ reach stationary values for $t \to \infty$, such that $[\eta_i(t)] = [\eta_i]$.

Under the same hypothesis, the second moment $[\eta_i(t)\eta_j(t+\tau)]$ is given by:

$$[\eta_i(t)\eta_j(t+\tau)] = \left[\sum_{k=1}^{N} J_{ik}\phi_k(t)\sum_{l=1}^{N} J_{jl}\phi_l(t+\tau)\right] + 2IJ(C_E - gC_I)[\phi] + I^2. \tag{2.26}$$

In order to evaluate the first term in the r.h.s., we differentiate two cases: first, we take $i = j$, yielding the noise auto-correlation. The sum over k (resp. l) can be split into a sum over k_E and k_I (resp. l_E and l_I) by segregating the contributions from the two populations. We thus get:

$$\begin{aligned}\left[\sum_{k=1}^{N} J_{ik}\phi_k(t)\sum_{l=1}^{N} J_{il}\phi_l(t+\tau)\right] &= \left[\sum_{k_E=1}^{N_E} J_{ik_E}\phi_{k_E}(t)\sum_{l_E=1}^{N_E} J_{il_E}\phi_{l_E}(t+\tau)\right] \\ + \left[\sum_{k_I=1}^{N_I} J_{ik_I}\phi_{k_I}(t)\sum_{l_I=1}^{N_I} J_{il_I}\phi_{l_I}(t+\tau)\right] &+ 2\left[\sum_{k_E=1}^{N_E} J_{ik_E}\phi_{k_E}(t)\sum_{l_I=1}^{N_I} J_{il_I}\phi_{l_I}(t+\tau)\right].\end{aligned} \tag{2.27}$$

We focus on the first term of the sum (same considerations hold for the second two), and we differentiate contributions from $k_E = l_E$ and $k_E \neq l_E$. Setting $k_E = l_E$ returns a contribution equal to $C_E J^2[\phi^2]$. In the sum with $k_E \neq l_E$, as C is fixed, we obtain exactly $C_E(C_E - 1)$ contributions of value $J^2[\phi]^2$. This gives, for the two populations:

$$\begin{aligned}&\left[\sum_{k=1}^{N} J_{ik}\phi_k(t)\sum_{l=1}^{N} J_{il}\phi_l(t+\tau)\right] = C_E J^2[\phi_i(t)\phi_i(t+\tau)] + C_E(C_E-1)J^2[\phi]^2 \\ &- 2C_E C_I g J^2[\phi]^2 + C_I g^2 J^2[\phi_i(t)\phi_i(t+\tau)] + C_I(C_I-1)g^2J^2[\phi]^2 \\ &= J^2(C_E + g^2C_I)[\phi_i(t)\phi_i(t+\tau)] + J^2(C_E - gC_I)^2[\phi]^2 - J^2(C_E + g^2C_I)[\phi]^2.\end{aligned} \tag{2.28}$$

By defining the rate auto-correlation function $C(\tau) = [\phi_i(t)\phi_i(t+\tau)]$, we finally get:

$$[\eta_i(t)\eta_i(t+\tau)] - [\eta_i]^2 = J^2(C_E + g^2C_I)\{C(\tau) - [\phi]^2\}. \tag{2.29}$$

When $i \neq j$, we instead obtain:

$$\left[\sum_{k=1}^{N} J_{ik}\phi_k(t)\sum_{l=1}^{N} J_{jl}\phi_l(t+\tau)\right] = C_E^2 J^2[\phi]^2 + pC_E J^2\{C(\tau) - [\phi]^2\} + C_I^2 g^2 J^2[\phi]^2 + pC_I g^2 J^2\{C(\tau) - [\phi]^2\} - 2C_E C_I g J^2[\phi]^2. \tag{2.30}$$

The constant p corresponds to the probability that, given that k is a pre-synaptic afferent of neuron i, the same neuron is connected also to neuron j. Because of sparsity, we expect this value to be small. More precisely, as N is assumed to be large, we can approximate the probability p with C/N. We eventually find:

$$[\eta_i(t)\eta_i(t+\tau)] - [\eta_i]^2 = pJ^2(C_E + g^2 C_I)\{C(\tau) - [\phi]^2\} \sim 0 \tag{2.31}$$

because $p \to 0$ when $N \to \infty$.

Once the statistics of the effective stochastic term η_i are known, we can describe the input current x in terms of its mean $\mu = [x_i]$ and its mean-subtracted correlation function $\Delta(\tau) = [x_i(t)x_i(t+\tau)] - [x_i]^2$. The mean field current $x_i(t)$ emerging from the stochastic process in Eq. 2.2 behaves as a time-correlated Gaussian variable. First we observe that, asymptotically, its mean value μ coincides with the mean of the noise term η_i:

$$\mu = J(C_E - gC_I)[\phi] + I. \tag{2.32}$$

By combining Eqs. 2.2 and 2.29, we moreover get the second equation for the auto-correlation evolution:

$$\ddot{\Delta}(\tau) = \Delta(\tau) - J^2(C_E + g^2 C_I)\{C(\tau) - [\phi]^2\}. \tag{2.33}$$

By explicitly constructing $x(t)$ and $x(t+\tau)$ in terms of unit Gaussian variables, we self-consistently rewrite the firing rate statistics $[\phi]$ and $C(\tau)$, as integrals over the Gaussian distributions:

$$\begin{aligned} [\phi] &= \int \mathcal{D}z\,\phi(\mu + \sqrt{\Delta_0}z) \\ C(\tau) &= \int \mathcal{D}z \left[\int \mathcal{D}x\,\phi(\mu + \sqrt{\Delta_0 - \Delta(\tau)}x + \sqrt{\Delta(\tau)}z)\right]^2. \end{aligned} \tag{2.34}$$

As we did in Section 2.1, we transform the problem into the equivalent classical motion in a one-dimensional potential. The potential $V(\Delta, \Delta_0)$ becomes:

$$V(\Delta, \Delta_0) = -\frac{\Delta^2}{2} + J^2(C_E + g^2 C_I)\left\{\int \mathcal{D}z \left[\int \mathcal{D}x\,\Phi(\mu + \sqrt{\Delta_0 - \Delta}x + \sqrt{\Delta}z)\right]^2 - \Delta[\phi]^2\right\} \tag{2.35}$$

where $\Phi(x)$ is the primitive of the threshold-linear activation function $\phi(x)$.

Similarly to the simpler model, the derivative of the potential in $\Delta = 0$ is always 0, suggesting a possible equilibrium point where the current distribution is concentrated in its mean value μ. Note that the existence of the stationary point in 0 stems from the $-\Delta[\phi]^2$ term in the potential, which comes from taking the connectivity degree C fixed for each unit in the

network (for a comparison with the equations obtained for random in-degree networks, see Chapter 3 and Appendix B).

Similarly again, the exact shape of the potential is controlled by the value of the synaptic strength J. A critical coupling J_C exists such that, for $J < J_C$, the only bounded solution is centered in μ and has $\Delta_0 = 0$. This solution corresponds to the homogeneous fixed point we analyzed in paragraph 7.3.4.

The chaotic solution for $J > J_C$ can be instead isolated by imposing:

$$V(\Delta_0, \Delta_0) = V(0, \Delta_0) \tag{2.36}$$

which transforms into:

$$\begin{aligned}\frac{\Delta_0^2}{2} =& J^2(C_E + g^2 C_I) \left\{ \int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0} z) - \left(\int \mathcal{D}z \Phi(\mu + \sqrt{\Delta_0} z) \right)^2 \right. \\ & \left. - \Delta_0 \left(\int \mathcal{D}z \phi(\mu + \sqrt{\Delta_0} z) \right)^2 \right\}.\end{aligned} \tag{2.37}$$

Note that the unknown first-order momentum μ enters explicitely in the equation for Δ_0, which has to be solved together with the equation for the mean:

$$\mu = J(C_E - gC_I) \int \mathcal{D}z \phi(\mu + \sqrt{\Delta_0} z) + I. \tag{2.38}$$

In a more compact form, we can reduce the system of equations to:

$$\begin{aligned}\mu &= J(C_E - gC_I)[\phi] + I \\ \frac{\Delta_0^2}{2} &= J^2(C_E + g^2 C_I) \left\{ [\Phi^2] - [\Phi]^2 - \Delta_0 [\phi]^2 \right\}.\end{aligned} \tag{2.39}$$

Once μ and Δ_0 are computed, their value can be injected into Eq. 2.33 to get the time course of the auto-correlation function.

We finally observe that, similarly to what we found in Section 2.2, the mean field equations admit a non-homogeneous stationary branch above J_C. Such solution, which corresponds to a chaotic regime in discrete-time models, is never stable for continuous-time dynamics, and will not be considered in most of the following analysis.

On self-averaging and E-I segregation Not surprisingly, the mean field equations we derived rely on the assumption of sparsity in the connectivity: $C \ll N$. Classic DMF theory, indeed, requires synaptic entries J_{ij} to be independent one from each other. Fixing the number of non-zero connections for each unit is imposing a strong dependence among the entries in each row of the synaptic matrix. Nevertheless, we expect this dependence to become very weak when $N \to \infty$, and we find that DMF can still predict correctly the system behavior, keeping however a trace of the network homogeneity through the term $-[\phi]^2$ in Eq. 2.33. Fixing the degree C sets to zero the asymptotic value of the auto-correlation function, and results in a perfect self-averaging and homogeneity of activity statistics in the population.

We remark that finding the DMF solution for an excitatory-inhibitory network reduces here to solving a system of two-equations. A large simplification in the problem comes here from considering networks where excitatory and inhibitory units receive statistically equivalent

inputs. DMF theory models indeed the statistical distribution of the input currents inside each network unit. For this reason, it does not include any element deriving from the segregation of the excitatory and the inhibitory populations in a two-columns connectivity structure. In consequence, for identical sets of parameters, we expect the same DMF equations to hold in more generic networks, where each neuron receive C_E excitatory and C_I inhibitory inputs, but can make excitatory or inhibitory output connections. In appendix A, we checked the validity of this observation.

In a more general case, where excitation and inhibition are characterized as distinguishable populations with their own statistics, solving the DMF equations becomes computationally costly. The main complication comes from the absence of any equivalent classical motion in a potential. For that reason, previous studies have focused mostly on the case of purely inhibitory populations [69, 58].

A systematic analysis of mean field equations and their solutions will be presented in the next chapter.

Two regimes of fluctuating activity 3

In this chapter, we solve the mean field equations for excitatory-inhibitory network models that we derived in detail in Chapter 2.

For fixed values of the activation upper-bound ϕ_{max}, we find that solutions are always finite, as the size of chaotic fluctuations is bounded. When the threshold-linear activation function has no saturation values, instead, the stability of the fluctuating activity regime depends on the exact value of the overall synaptic coupling. To begin with, we rigorously compute the boundary between stable and run-away activity regions for the simplest network architecture where there are no external inputs and the connectivity in-degree C is fixed.

In a second step, we check how different biological constraints (details on the network architecture, spiking noise etc.) qualitatively and quantitatively perturb the dynamical regimes of the network. We find that stable chaotic activity, where fluctuations are bounded solely by recurrent inhibition, robustly appears in many generalized network models.

3.1 Dynamical Mean Field solutions

In all the cases that we consider, extracting solutions from the DMF equations corresponds to solving a system of several non-linear coupled equations which involve single or multiple Gaussian integrals. Importantly, for threshold-linear activation functions, an analytical expression for many of these integrals can be derived. Iteration is a practical and stable method for exactly computing the solutions of the DMF system of equations.

For a network architecture where the number of incoming connections is fixed, DMF theory reduces to a system of two equations (Eq. 2.39).

In agreement with the dynamical systems analysis in paragraph 7.3.4, at low coupling values, the DMF theory predicts a solution with vanishing variance and auto-correlation (Fig. 3.1 **a**). Input currents set into a stationary and uniform value, corresponding to their mean μ. The predicted value of μ coincides with the homogeneous fixed point x_0, representing a low firing rate background activity (Fig. 3.1 **c**). As the coupling J is increased, the mean current becomes increasingly negative because inhibition dominates, and the mean firing rate decreases (Fig. 3.1 **c-d**).

For a critical coupling strength $J = J_C$ (which coincides with J_0, where the fixed point

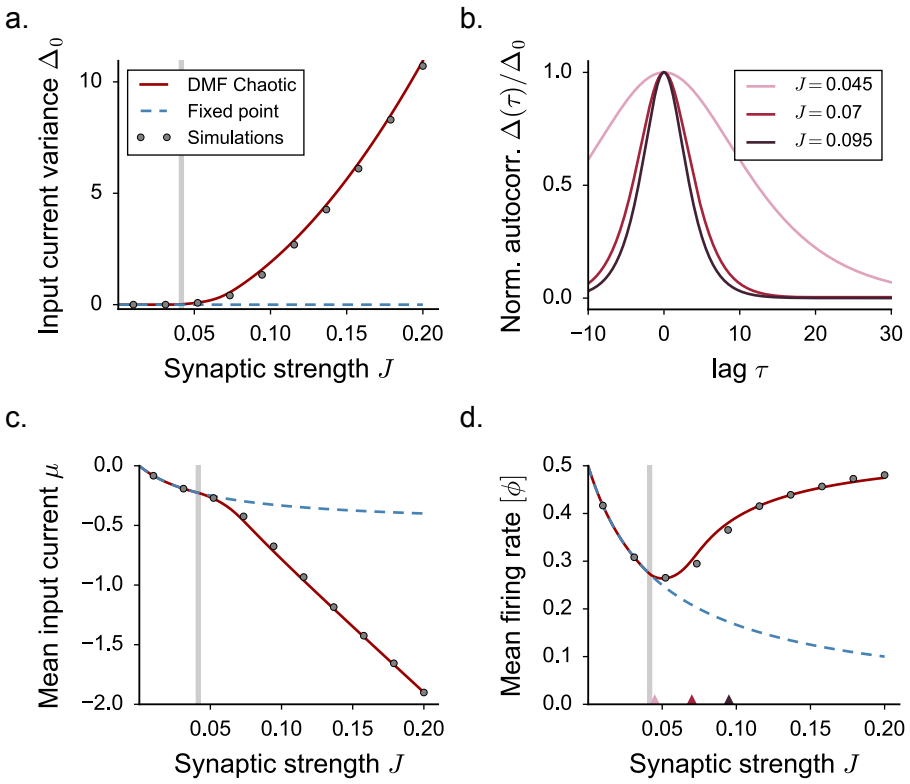


Figure 3.1: DMF description of network activity with a threshold-linear activation function. The dynamics mean field results are shown in full lines, numerical simulations as points. **a.** Input current variance as a function of the synaptic coupling J. Vertical grey lines indicate the critical value J_C. Grey points show time and population averages performed on 4 realizations of simulated networks, $N = 7000$. **b.** Normalized auto-correlation function for increasing values of the synaptic coupling (indicated by colored triangles in panel **d**). **c-d.** First order statistics: mean input current and mean firing rate. Choice of the parameters: $g = 5, C = 100$, $\phi_{max} = 2$.

solution loses stability), DMF predicts the onset of a second solution with fluctuations of non-vanishing magnitude. Above J_C, the variance of the activity grows smoothly from zero (Fig. 3.1 **a**), and the auto-correlation $\Delta(\tau)$ acquires a temporal structure, exponentially decaying to zero as $\tau \to \infty$. Close to the critical coupling, the dynamics exhibit a critical slowing down and the decay timescale diverges at J_C, a behavior characteristic of a critical phase transition [127] (Fig. 3.1 **b**).

The onset of irregular, fluctuating activity is characterized by a transition of the second-order statistics from zero to a non-vanishing value. The appearance of fluctuations, however, directly affects also the first-order statistics. As the firing rates are constrained to be positive, large fluctuations induce deviations of the mean firing rate $[\phi]$ and the mean input current μ from their fixed point solutions. In particular, as J increases, larger and larger fluctuations in the current lead to an effective increase in the mean firing rate although the network is inhibition-dominated (Fig. 3.1 **a-c-d**). The increase in mean firing rate with synaptic coupling is therefore a signature of the onset of fluctuating activity in this class of excitatory-inhibitory

networks.

3.2 Intermediate and strong coupling chaotic regimes

The mean field approach revealed that, above the critical coupling J_C, the network generates fluctuating but stable, stationary activity. The dynamical systems analysis, however, showed that the dynamics of an equivalent linearized network are unstable and divergent for identical parameter values. The stability of the fluctuating activity is therefore necessarily due to the two non-linear constraints present in the system: the requirement that firing rates are positive, and the requirement that firing rates are limited by an upper bound ϕ_{max}.

In order to isolate the two contributions, we examined how the amplitude of fluctuating activity depends on the upper bound on firing rates ϕ_{max}. Ultimately, we take this bound to infinity, leaving the activity unbounded. Solving the corresponding DMF equations revealed the presence of two qualitatively different regimes of fluctuating activity above J_c (Fig. 3.2).

For intermediate coupling values, the magnitude of fluctuations and the mean firing rate depend only weakly on the upper bound ϕ_{max}. In particular, for $\phi_{max} \to \infty$ the dynamics remain stable and bounded. The positive feedback that generates the linear instability is dominantly due to negative, inhibitory interactions multiplying positive firing rates in the linearized model. In this regime, the requirement that firing rates are positive, combined with dominant inhibition, is sufficient to stabilize this feedback and the fluctuating dynamics.

For larger coupling values, the dynamics depend strongly on the upper bound ϕ_{max}. As ϕ_{max} is increased, the magnitude of fluctuations and the mean firing rate continuously increase and diverge for $\phi_{max} \to \infty$. For large coupling values, the fluctuating dynamics are therefore stabilized by the upper bound and become unstable in absence of saturation, even though inhibition is globally stronger than excitation.

Fig. 3.2 **d** summarizes the qualitative changes in the dependence on the upper bound ϕ_{max}. In the fixed point regime, mean inputs are suppressed by inhibition, and they correspond to the low-gain region of $\phi(x)$, which is independent of ϕ_{max}. Above J_C, in the intermediate regime, the solution rapidly saturates to a limiting value. In the strong coupling regime, the mean firing rate, as well as the mean input μ, and its standard deviation $\sqrt{\Delta_0}$ grow linearly with the upper bound ϕ_{max}. We observe that when ϕ_{max} is large, numerically simulated mean activity show larger deviations from the theoretically predicted value, because of larger finite size effects (for a more detailed discussion, see in Appendix A).

The two regimes of fluctuating activity are characterized by different scalings of the first- and second-order statistics with the upper-bound ϕ_{max}. In the absence of upper bound on the activity, i.e. in the limit $\phi_{max} \to \infty$, the two regimes are sharply separated by a second "critical" coupling J_D: below J_D, the network reaches a stable fluctuating steady-state and DMF admits a solution; above J_D, the network has no stable steady-state, and DMF admits no solution. J_D corresponds to the value of the coupling for which the DMF solution diverges, and can be determined analytically (see in next paragraph). For a fixed, finite value of the upper bound ϕ_{max}, there is however no sign of transition as the coupling is increased past J_D. Indeed, for a fixed ϕ_{max}, the network reaches a stable fluctuating steady state on both sides of J_D, and no qualitative difference is apparent between these two steady states. The difference appears only when the value of the upper bound ϕ_{max} is varied. J_D therefore separates two dynamical regimes in which the statistics of the activity scale differently with the upper-bound ϕ_{max}, but for a fixed, finite ϕ_{max} it does not correspond to an instability. The second "critical" coupling

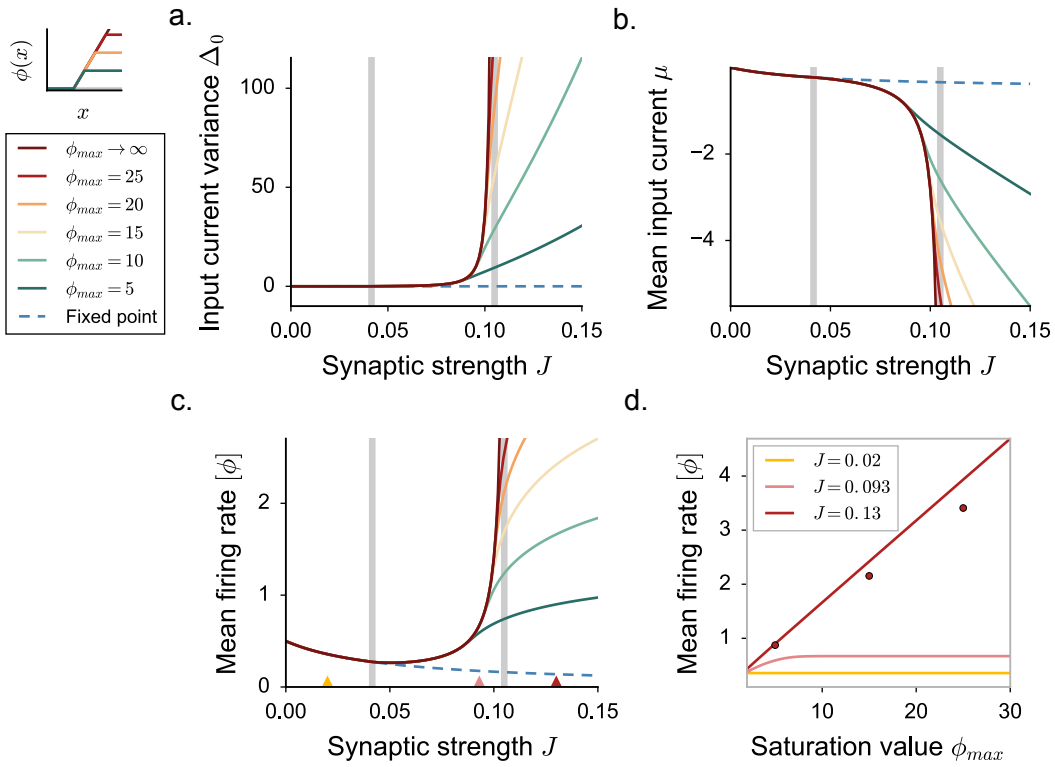


Figure 3.2: Appearance of three dynamical regimes in excitatory-inhibitory rate networks, dynamical mean field predictions. Threshold-linear activation function saturating at different values of the upper bound ϕ_{max}. **a-b-c.** DMF characterization of the statistics for different values of the saturation value ϕ_{max}. In **a**, input current variance, in **b**, input current mean, in **c**, mean firing rate. Vertical grey lines indicate the critical couplings J_C and J_D. **d.** Mean firing rate dependence on the upper bound ϕ_{max}, for three coupling values corresponding to the three different dynamical regimes (indicated by triangles in panel **c**). Dots show time and population averages performed on 4 realizations of simulated networks, $N = 6000$. Choice of the parameters: $g = 5$, $C = 100$.

J_D is therefore qualitatively different from the critical coupling J_C, which is associated with an instability for any value of ϕ_{max}.

In summary, the two non-linearities induced by the two requirements that the firing rates are positive and bounded play asymmetrical roles in stabilizing fluctuating dynamics. In excitatory-inhibitory networks considered here, this asymmetry leads to two qualitatively different fluctuating regimes.

3.2.1 Computing J_D

In order to obtain a closed expression for computing J_D, we study the behavior of the DMF solution close to the second critical coupling J_D, for a non-saturating activation function where $\phi_{max} \to \infty$. When J approaches J_D, $\Delta_0 \to \infty$, while $\mu \to -\infty$.

Led by dimensionality arguments, we assume that, close to the divergence point, the ratio

$k = \mu/\sqrt{\Delta_0}$ is constant. With a threshold-linear transfer function, it is possible to compute analytically the three Gaussian integrals implicit in Eq. 2.39 and to provide an explicit analytic form of the DMF equations. The equation for the mean translates into:

$$\mu = J(C_E - gC_I)[\phi] = J(C_E - gC_I)\left\{\left(\frac{1}{2}+\mu\right)\left(\frac{1}{2}-g(x_a)\right)+\sqrt{\frac{\Delta_0}{2\pi}}e^{-\frac{1}{2}x_a^2}\right\} \quad (3.1)$$

where $x_a = \frac{1}{\sqrt{\Delta_0}}\left(-\frac{1}{2}-\mu\right) \sim -k$ and where we have defined: $g(x) = \frac{1}{2}\,\mathrm{erf}(x/\sqrt{2})$. When $J \to J_D$, by keeping only the leading order in $\sqrt{\Delta_0}$, we find $\mu = \hat{k}\sqrt{\Delta_0}$ with:

$$\hat{k} = \frac{J(C_E - gC_I)\frac{e^{-\frac{k^2}{2}}}{\sqrt{2\pi}}}{1 - J(C_E - gC_I)(\frac{1}{2}+G(k))}. \quad (3.2)$$

By imposing $k = \hat{k}$, one can determine self-consistently the value of k for each value of J. We introduce $\mu = k\sqrt{\Delta_0}$ into the second equation for Δ_0. By keeping only the leading order in Δ_0, we find:

$$\begin{aligned}\sqrt{\Delta_0} &= f(k)\\ f(k) &= \frac{J^2(C_E + g^2C_I)T(k)}{\frac{1}{2} - J^2(C_E + g^2C_I)S(k)}\end{aligned} \quad (3.3)$$

with:

$$\begin{aligned}S(k) =& \frac{1}{4}k^4\left[\frac{1}{2}+g(k)\right]+\frac{1}{4}k^3\frac{e^{-\frac{k^2}{2}}}{\sqrt{2\pi}}+k^2\left[\frac{3}{2}\left(\frac{1}{2}+g(k)\right)-\left(\frac{1}{2}+g(k)\right)^2\right]\\ &+k\left[\frac{5}{4}\frac{e^{-\frac{k^2}{2}}}{\sqrt{2\pi}}-2\left(\frac{1}{2}+g(k)\right)\frac{e^{-\frac{k^2}{2}}}{\sqrt{2\pi}}\right]+\frac{3}{4}\left(\frac{1}{2}+g(k)\right)-\left(\frac{e^{-\frac{k^2}{2}}}{\sqrt{2\pi}}\right)^2\\ &-\left\{\left(\frac{1}{2}k^2+\frac{1}{2}\right)\left[\frac{1}{2}+g(k)\right]+\frac{1}{2}k\frac{e^{-\frac{k^2}{2}}}{\sqrt{2\pi}}\right\}^2.\end{aligned} \quad (3.4)$$

In order to obtain a solution Δ_0, from Eq. 3.3 we require the function $f(k)$ to be positive. We observe that f diverges when its denominator crosses zero. Here $f(k)$ changes sign, becoming negative. We use this condition to determine J_D (Fig. 3.3):

$$J_D^2(C_E + g^2C_I)S(k(J_D)) = \frac{1}{2}. \quad (3.5)$$

The value of J_D depends both on the relative strength of inhibition g, and the total number of incoming connections C (Fig. 3.3). Increasing either g or C increases the total variance of the interaction matrix J_{ij}, shifting the instability of the homogeneous fixed point to lower couplings. The size of the intermediate fluctuating regime however depends only weakly on the number of incoming connections C (Fig. 3.3 **a**). In contrast, increasing the relative strength of inhibition diminishes the influence of the upper bound and enlarges the phase space region corresponding to the intermediate regime, where fluctuations are stabilized intrinsically

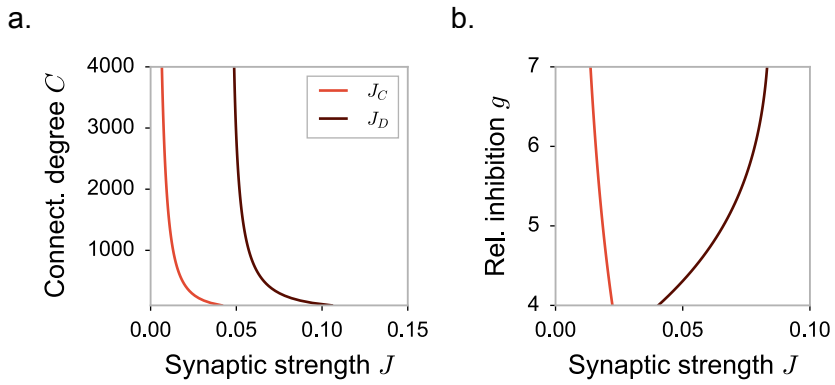


Figure 3.3: Phase diagram of the dynamics: dependence on the connectivity in-degree C (**a**) and on the inhibition dominance parameter g (**b**). All other parameters are kept fixed as in Fig. 3.2.

by recurrent inhibition (Fig. 3.3 **b**). The second critical coupling J_D is in particular expected to increase with g and diverge for purely inhibitory networks. However, for very large relative inhibition, numerical simulations show strong deviations from DMF predictions, due to the breakdown of the Gaussian approximation which overestimates positive feedback (see Appendix A).

3.2.2 Purely inhibitory networks

To identify the specific role of excitation in the dynamics described above, we briefly consider here the case of networks consisting of a single inhibitory population. Purely inhibitory networks display a transition from a fixed point regime to chaotic fluctuations [69, 58]. The amplitude of fluctuations appears to be in general much smaller than in excitatory-inhibitory networks, but increases with the constant external current I (Fig. 3.4 **a**). In contrast to our findings for networks in which both excitation and inhibition are present, in purely inhibitory networks intrinsically generated fluctuations lead to a very weak increase in mean firing rates compared to the fixed point (Fig. 3.4 **b-c**). This effect can be understood by noting that within the dynamical mean field theory, the mean rate is given by $(\mu - I)/J(C_E - gC_I)$. The term $C_E - gC_I$ in the denominator determines the sensitivity of the mean firing rate to changes in mean input. This term is always negative as we are considering inhibition-dominated networks, but its absolute value is much smaller in presence of excitation, i.e. when excitation and inhibition approximately balance, compared to purely inhibitory networks. As the onset of intrinsically generated fluctuations modifies the value of the mean input with respect to its value in the fixed point solution (Fig. 3.1 **c**, Fig. 3.4 **b**), this simplified argument explains why the mean firing rates in the inhibitory network are much less sensitive to fluctuations than in the excitatory-inhibitory case.

Moreover, the second fluctuating regime found in E-I networks does not appear in purely inhibitory networks. Indeed, the divergence of first- and second-order statistics that occurs in E-I networks requires positive feedback that is absent in purely inhibitory networks. Note that for purely inhibitory, sparse networks, important deviations can exist at very large couplings between the dynamical mean field theory and simulations (see Appendix A for a more detailed

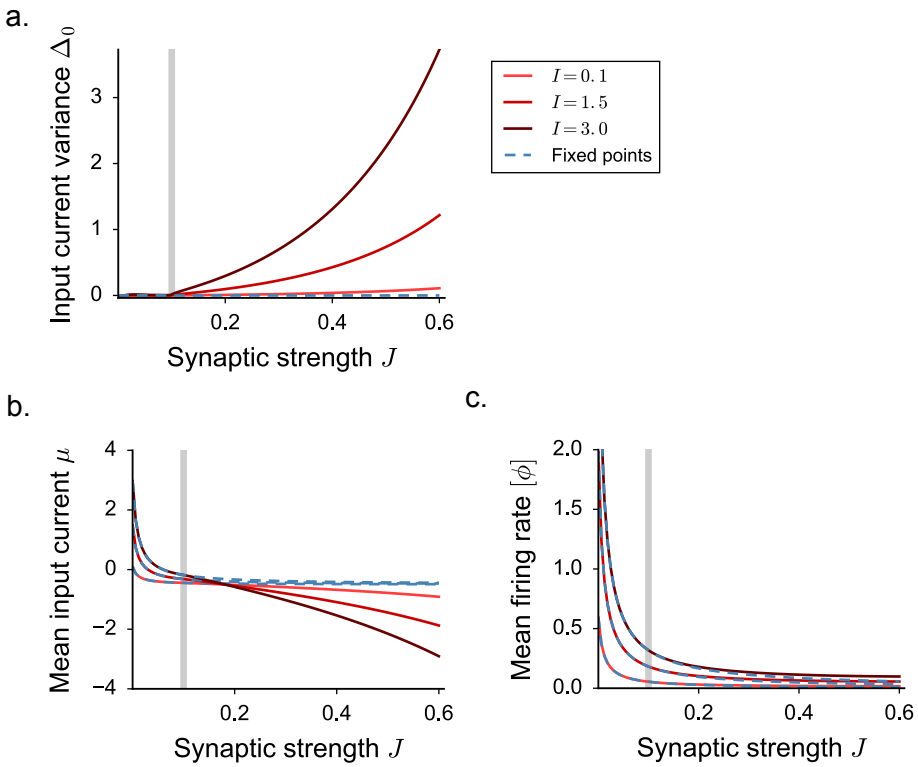


FIGURE 3.4: Statistical description of the activity in purely inhibitory networks. Results of the dynamical mean field theory (obtained through setting $C_E = 0$ and $g = 1$) for different values of the excitatory external current I. **a.** Input current variance, **b.** mean current and **c.** mean firing rate as a function of the synaptic coupling J. Vertical grey lines indicate the critical value J_C.

discussion).

The two main findings reported above, the strong influence of intrinsically generated fluctuations on mean firing rate, and the existence of two different fluctuating regimes therefore critically rely on the presence of excitation in the network.

3.3 Extensions to more general classes of networks

In a second step, we extend our analysis to more complex models of excitatory-inhibitory networks. In all the cases that we study, the DMF equations can still be derived and solved numerically, but an analytical expression for the divergence coupling J_D is typically harder to derive.

In Appendix B we show in details how the mean field equations should be modified in order to include the new additional constraints within the self-consistent description. Here, we focus on the results and their implications.

3.3.1 The effect of noise

To begin with, we investigated whether the two different fluctuating regimes described above can be still observed when spiking noise is added to the dynamics. Following [69], we added a Poisson spiking mechanism on the rate dynamics in Eq. 2.21, and let the different units interact through spikes (see Appendix B). Within a mean field approach, interaction through spikes lead to an additive white noise term in the dynamics [69, 55]. To determine the effect of this additional term on the dynamics, we first treated it as external noise and systematically varied its amplitude as a free parameter.

The main effect of noise is to induce fluctuations in the activity for all values of network parameters (Fig. 3.5 **a**). As a result, in presence of noise, the sharp transition between constant and fluctuating activity is clearly lost. The feedback mechanism that generates intrinsic fluctuations nevertheless still operates and strongly amplifies the fluctuations induced by external noise.

The DMF framework can be extended to include external noise and determine the additional variability generated by network feedback (see also Appendix B). When the coupling J is small, the temporal fluctuations in the activity are essentially generated by the filtering of external noise. Beyond the original transition at J_C, instead, when the feedback fluctuations grow rapidly with synaptic coupling, the contribution of external noise becomes rapidly negligible with respect to the intrinsically-generated fluctuations (Fig. 3.5 **a**).

As shown also in other studies [69, 55], a dramatic effect of introducing external noise is a strong reduction of the timescale of fluctuations close to J_C. In absence of noise, just above the fixed point instability at J_C, purely deterministic rate networks are characterized by the onset of infinitely slow fluctuations. These slow fluctuations are however of vanishingly small magnitude, and strongly sensitive to external noise. Any finite amount of external noise eliminates the diverging timescale. For weak external noise, a maximum in the timescale can be still seen close to J_C, but it quickly disappears as the magnitude of noise is increased. For modest amounts of external noise, the timescale of the fluctuating dynamics becomes a monotonic function of synaptic coupling (Fig. 3.5 **b**).

While in presence of external noise there is therefore no formal critical phase transition, the dynamics still smoothly change from externally-generated fluctuations around a fixed point into intrinsically-generated, non-linear fluctuations. This change of regime is not necessarily reflected in the timescale of the dynamics, but can clearly be seen in the excess variance, and also in the first-order statistics such as the mean-firing rate, which again strongly increases with coupling. Moreover, we found that the existence of the second fluctuating regime is totally insensitive to noise: above the second critical coupling J_D, the activity is only stabilized by the upper bound on the firing rates, and diverges in its absence. In that parameter region, intrinsically-generated fluctuations diverge, and the external noise contributes only a negligible amount.

We considered so far the effect of an external white noise of arbitrary amplitude. If that noise represents spiking interactions, its variance is however not a free parameter, but instead given by $J^2(C_E + g^2 C_I)[\phi]/\bar{\tau}$. In particular, the amplitude of spiking noise increases both with the synaptic coupling and with the mean firing rate $[\phi]$, which itself depends on the coupling and fluctuations as pointed out above. As a result, the amplitude of the spiking noise dramatically increases in the fluctuating regime (Fig. 3.5 **d**). When J becomes close to the second critical coupling J_D, the spiking noise however still contributes only weakly to the total variance (see in Appendix B), and the value of J_D is not affected by it (Fig. 3.6 **a-b**). The

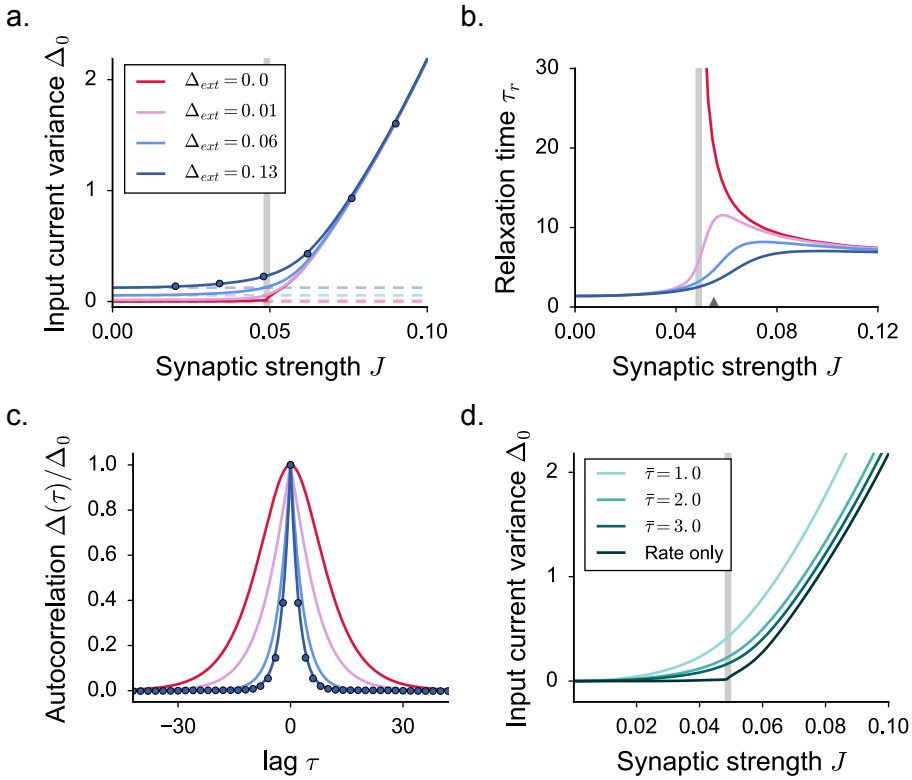


FIGURE 3.5: Statistical description of the activity in excitatory-inhibitory networks with external and spiking noise. The dynamical mean field results are shown in full lines, numerical simulations as points. **a.** Input current variance in presence of external noise, for increasing values of the noise amplitude (white noise, variance equal to $2\Delta_{ext}$). Blue dots: results of numerical simulations for $\Delta_{ext} = 0.13$, $N = 7500$, average of 4 realizations of the synaptic matrix. The grey vertical line shows the critical coupling J_C in the deterministic model. Dashed lines indicate the statistics of an effective fixed point, where the only variance is generated by the noise contribution Δ_{ext}. The fixed point firing rate is computed as a Gaussian average, with the mean given by the fixed point x_0 and the variance provided solely by the noise term. The deflection from the effective fixed point underlines an internal amplification of noise produced by network feedback. **b.** Fluctuations relaxation time, measured as the auto-correlation $\Delta(\tau)$ full width at half maximum. **c.** Normalized auto-correlation for fixed J and different levels of noise. The corresponding coupling value is indicated on the x axis of panel **b**. **d.** Input variance in a network with spiking dynamics, where spikes are generated according to inhomogeneous Poisson processes. Increasing the time constant of rate dynamics $\bar{\tau}$ (see in Appendix B) decreases the amplitude of spiking noise. Choice of the parameters: $g = 4.1$, $C = 100$.

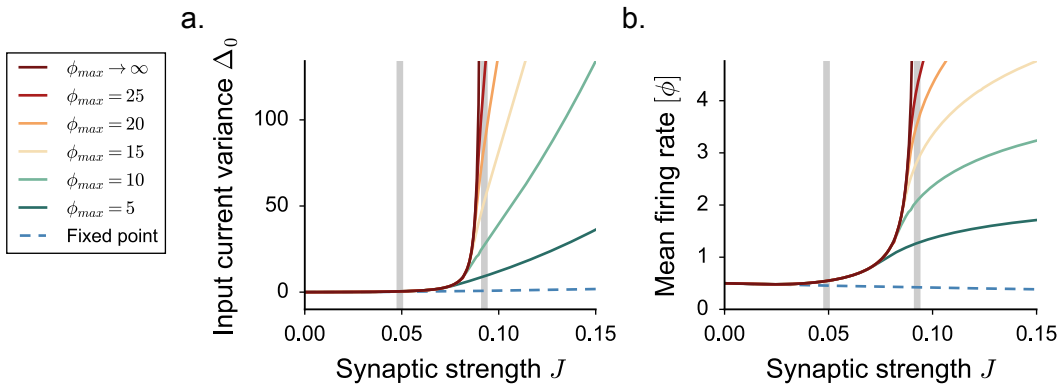


Figure 3.6: Appearance of the three dynamical regimes in a network with spiking noise: input current variance (**a**) and mean firing rate (**b**) for different saturation values ϕ_{max}. Choice of the parameters: $g = 4.1$, $C = 100$.

amplitude of spiking noise is also inversely proportional the timescale $\bar{\tau}$ of the dynamics (see Eq. 6 in Appendix B). Slower dynamics tend to smooth out fluctuations due to spiking inputs (Fig. 3.5 **d**), reduce the amount of spiking and noise and therefore favor the appearance of slow fluctuations close to the critical coupling J_C [69].

In conclusion, the main new findings reported above, the influence of intrinsically generated fluctuations on mean firing rate, and the existence of two different fluctuating regimes are still observed in presence of external or spike-generated noise. In particular, above the second transition, intrinsically generated fluctuations can be arbitrarily strong and therefore play the dominant role with respect to external or spiking noise.

3.3.2 Connectivity with stochastic in-degree

We now turn to networks in which the number of incoming connections is not fixed for all the neurons, but fluctuates stochastically around a mean value C. We consider a connectivity scheme in which each excitatory (resp. inhibitory) neuron makes a connection of strength J (resp. $-gJ$) with probability C/N.

In this class of networks, the number of incoming connections per neuron has a variance equal to the mean. As a consequence, in the stationary state, the total input strongly varies among units. In contrast to the case of a fixed in-degree, the network does not admit an homogeneous fixed point. The fixed point is instead heterogeneous, and more difficult to study using dynamical systems tools.

The dynamical mean field approach can however be extended to include the heterogeneity generated by the variable number of incoming connections [141, 69, 58]. As derived in Appendix B, the stationary distributions are now described by a mean and a static variance Δ_0 that obey:

$$\begin{aligned} \mu &= J(C_E - gC_I)[\phi] + I, \\ \Delta_0 &= J^2(C_E + g^2 C_I)[\phi^2]. \end{aligned} \tag{3.6}$$

The stationary solution loses stability at a critical value $J = J_C$. In the strong coupling regimes, DMF predicts the onset of a time-dependent solution with a decaying autocorrelation

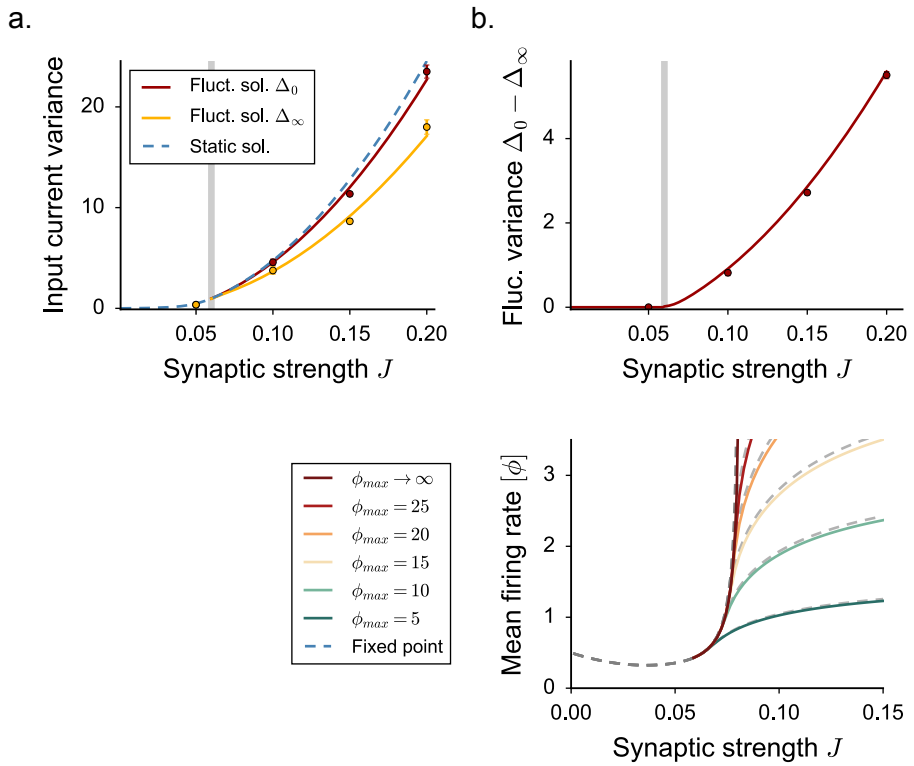


FIGURE 3.7: Mean field characterization of the activity in networks with stochastic in-degree. The dynamical mean field results are shown in full lines, numerical simulations as points. **a.** Total input current variance Δ_0. The heterogeneity in the connectivity induces an additional quenched variance Δ_∞ (shown in dashed blue for the fixed point, and yellow for the fluctuating solution, where it corresponds to Δ_0). Red (resp. yellow) points show time and population averages of Δ_0 (resp. Δ_∞) performed on 3 realizations of simulated networks, $N = 6500$. **b.** Isolated contribution of temporal fluctuations to the variance. **(c)** Mean firing rate, for different values of the saturation ϕ_{max}. Grey dashed lines indicate the stationary solution, becoming a thick colored line, corresponding to the chaotic phase, at J_C. Choice of the parameters: $g = 5$, $C = 100$, $\phi_{max} = 2$.

function, with initial condition Δ_0 and asymptotic value Δ_∞. The values of μ, Δ_0 and Δ_∞ are determined as solution of a system of three equations (see in Appendix B). In this regime, the effective amplitude of temporal fluctuations is given by the difference $\Delta_0 - \Delta_\infty$ (Fig. 3.7 **b**). A non-zero value of Δ_∞ reflects the heterogeneity in the connectivity and indicates a qualitative change in the dynamics: single neuron activity is not ergodic and stays highly self-correlated even after long times. Note moreover that because the static variance increases with coupling (Fig. 3.7 **a**), the mean activity increases with coupling for the static solution. In the fluctuating regime, as the additional temporal variance $\Delta_0 - \Delta_\infty$ is weaker than the static variance Δ_∞, temporal fluctuations do not lead to a noticeable increase in mean firing rate with respect to the static solution (Fig. 3.7 **c**).

Fig. 3.7 **c** displays the dependence on the upper bound ϕ_{max} of the mean field solution. We first note that in networks with very variable in-degree , the critical value J_C weakly de-

pends on the saturation upper bound due to large static heterogeneity. Above J_C, an intermediate regime exists where the activity is stabilized by inhibition, and remains finite even in absence of upper bound. For couplings above a second critical coupling J_D, the dynamics are stabilized only by the upper bound ϕ_{max}. Networks with variable in-degree therefore show the same three dynamical regimes as networks with fixed degree, the main difference being that variable in-degree can reduce the extent of the intermediate regime between J_C and J_D.

3.3.3 General excitatory-inhibitory networks

In the class of networks we investigated so far, excitatory and inhibitory units received statistically equivalent inputs. Under this assumption, the network dynamics are characterized by a single mean and variance for both excitatory and inhibitory populations, which considerably simplifies the mean field description. Here we relax this assumption and show that the properties of intrinsically generated fluctuations described so far do not critically depend on it.

We consider a more general class of networks, in which synaptic connections are arranged in a block matrix:

$$\mathrm{J} = J \left(\begin{array}{c|c} \mathrm{J_{EE}} & \mathrm{J_{EI}} \\ \hline \mathrm{J_{IE}} & \mathrm{J_{II}} \end{array} \right) \tag{3.7}$$

where each block $\mathrm{J_{kk'}}$ is a sparse matrix, containing on each row $C_{kk'}$ non-zero entries of value $j_{kk'}$. The parameter J represents a global scaling on the intensity of the synaptic strength. For the sake of simplicity, we restrict ourselves to the following configuration: each row of J contains exactly C_E non-zero excitatory entries in the blocks of the excitatory column, and exactly C_I inhibitory entries in the inhibitory blocks. Non-zero elements are equal to j_E in $\mathrm{J_{EE}}$, to $-g_E j_E$ in $\mathrm{J_{EI}}$, to j_I in $\mathrm{J_{IE}}$, and to $-g_I j_I$ in $\mathrm{J_{II}}$. The previous case is recovered by setting $j_E = j_I = 1$ and $g_E = g_I$.

The network admits a fixed point in which the activities are different for excitatory and inhibitory units, but homogeneous within the two populations. This fixed point is given by:

$$\begin{pmatrix} x_0^E \\ x_0^I \end{pmatrix} = J \begin{pmatrix} j_E(C_E\phi(x_0^E) - g_E C_I \phi(x_0^I)) \\ j_I(C_E\phi(x_0^E) - g_I C_I \phi(x_0^I)) \end{pmatrix} \tag{3.8}$$

where x_0^E and x_0^I are the fixed-point inputs to the two populations.

The linear stability of the fixed point is determined by the eigenvalues of the matrix:

$$\mathrm{S} = J \left(\begin{array}{c|c} \phi'(x_0^E)\mathrm{J_{EE}} & \phi'(x_0^I)\mathrm{J_{EI}} \\ \hline \phi'(x_0^E)\mathrm{J_{IE}} & \phi'(x_0^I)\mathrm{J_{II}} \end{array} \right) \tag{3.9}$$

The fixed point is stable if the real part of all the eigenvalues is smaller than one. As for simple, column-like E-I matrices, the eigenspectrum of S is composed of a discrete and a densely distributed part, in which the bulk of the eigenvalues are distributed on a circle in the complex plane [6, 7, 5]. The discrete component consists instead of two eigenvalues, which in general can be complex, potentially inducing various kinds of fixed point instabilities (for the details, see Appendix B). As in the previous paragraphs, we consider a regime where both g_E and g_I are strong enough to dominate excitation, and the outlier eigenvalues have negative real part. In those conditions, the first instability to occur is the chaotic one, where the radius of the complex circle of the eigenspectrum crosses unity. This radius increases with the overall coupling J, defining a critical value J_C where the fixed point loses stability.

Dynamical mean field equations for the fluctuating regime above the instability are, in this general case, much harder to solve as they now involve two means and two auto-correlation functions, one for each populations. For that reason, we restrict ourselves to a slightly different dynamical system with discrete-time evolution:

$$x_i(t+1) = \sum_{j=1}^{N} J_{ij}\phi(x_j(t)). \tag{3.10}$$

Such a network corresponds to extremely fast dynamics with no current filtering (Fig. 3.8 **a-b**). Previous works [89, 30, 141] have studied that class of models in case of synaptic matrices that lacked E-I separation, and for activation functions that were symmetric. These works pointed out strong analogies with the dynamics emerging in continuous time [127]. Discrete-time dynamics can however induce a new, period-doubling bifurcation when inhibition is strong. We therefore restrict the analysis to a regime where inhibition is dominating but not excessively strong. Notice that in general, outside the range of parameters considered in this analysis, we expect generic E-I networks to display a richer variety of dynamical regimes.

To begin with, we observe that the fixed-point (Eq. 3.8) and its stability conditions (Eq. 3.9) are identical for continuous and discrete dynamics. For discrete time, the DMF equations are however much simpler than for continuous dynamics, and can be easily fully solved even if the two populations are characterized now by different values of mean and variance.

Solving the DMF equations confirms that the transition to chaos in this class of models is characterized by the same qualitative features as before (Fig. 3.8 **c-d**). As the order parameter J is increased, the means and the variances of both the E and the I population display a transition from the fixed point solution to a fluctuating regime characterized by positive variance Δ_0 and increasing mean firing rate. By smoothly increasing the upper bound of the saturation function ϕ_{max} as before, we find a second critical value J_D at which the firing activity of both populations diverge (Fig. 3.8 **e-f**). We conclude that the distinction in three regimes reported so far can be extended to discrete-time dynamics; in this simplified framework, our results extend to more general E-I connectivity matrices.

3.4 Relation to previous works

The transition from fixed point to fluctuating activity was first studied by Sompolinsky, Crisanti and Sommers [127]. As discussed in details in Chapter 2, in that classical work, the connectivity was Gaussian and the activation function symmetric around zero, so that the dynamics exhibited a sign-reversal symmetry. An important consequence of this symmetry is that the mean activity was always zero, and the transition was characterized solely in terms of second-order statistics, which were described through a dynamical mean field equation.

Recent studies have examined more general and biologically plausible networks [6, 7, 58, 69]. Two of those studies [58, 69] derived dynamical mean field (DMF) equations to networks with segregated excitatory and inhibitory populations, and asymmetric, positively defined transfer functions. The DMF equations are however challenging to solve in the general case of two distinct excitatory and inhibitory populations (see Appendix B). The two studies therefore analyzed in detail DMF solutions for purely inhibitory networks, and explored fluctuating activity in excitatory-inhibitory networks mainly through simulations.

In contrast to these recent works, here we exploited a simplified network architecture, in which DMF equations can be fully analyzed for excitatory-inhibitory networks. We found the

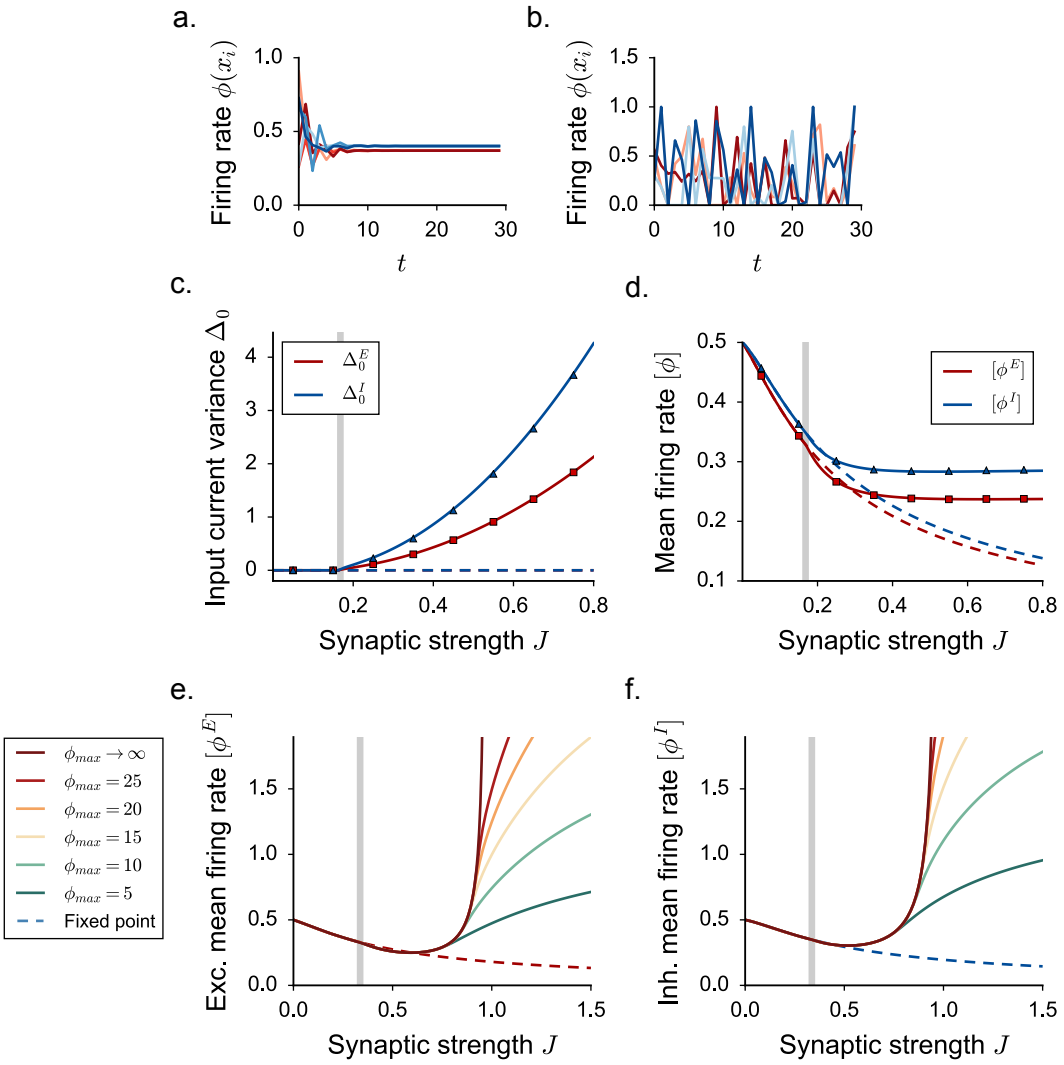


Figure 3.8: Fluctuating dynamics in more general networks where excitatory and inhibitory neurons are not statistically equivalent. Discrete-time rate evolution. **a-b.** Network discrete-time activity: numerical integration of the Eq. 3.10, firing rates of randomly selected units. Excitatory neurons are plotted in the red scale, inhibitory ones in the blue one. $N = 1000$. In **a**, $J < J_C$; in **b**, $J > J_C$. **c-d.** Statistical characterization of network activity, respectively in terms of the input variance and the mean firing rate. Dynamical mean field results are shown in full lines. Dashed lines: fixed points. Dots: numerical simulations, $N = 7500$, average over 3 realizations. Vertical grey lines indicate the critical value J_C. $\phi_{max} = 1$. **e-f.** Mean firing rate for different values of the saturation ϕ_{max}, in the excitatory and the inhibitory population. Choice of the parameters: $j_E = 0.1$, $j_I = 1.5 j_E$, $g_E = 4.5$, $g_I = 4.2$, $C = 100$.

presence of excitation qualitatively changes the nature of the dynamics, even though inhibition dominates. In purely inhibitory networks, fluctuations are weaker than in excitatory-inhibitory networks, and as a result do not affect first-order statistics.

In [69], the authors used transfer functions without upper bounds, and found that the chaotic state can undergo an instability in which the activity diverges. This instability is directly related to the transition between the two fluctuating regimes which we studied in detail for bounded transfer functions. Here we showed that these two dynamical regimes can in fact be distinguished only if the upper bound is varied: for a fixed upper bound, there is no sign of a transition. Moreover, we showed that excitation is required for the appearance of the second fluctuating regime, as this regime relies on positive feedback. For purely inhibitory networks, in which positive feedback is absent, simulations show that the second fluctuating regime does not occur, although it is predicted by dynamical mean field theory: indeed DMF relies on a gaussian approximation which does not restrict the interactions to be strictly negative, and therefore artifactually introduces positive feedback at strong coupling.

Finally, previous studies [58, 69] focused on networks with random in-degree or Gaussian coupling. In such networks, the quenched component of the coupling matrix leads to quenched heterogeneity in the stationary solution. In the present work, we instead mostly studied networks with fixed in-degree. We showed that in such a setting a homogeneous distribution is the stable solution, so that the quenched variability is not required for the transition to fluctuating activity.

Rate fluctuations in spiking networks 4

In the spirit of investigating how the classical results from [127] apply to realistic models of cortical circuits, we studied the dynamics of constrained networks of firing rate units. Up to this point, we did not specify how rate units should be interpreted in terms of single cortical neurons, whose activity is characterized instead in terms of discrete action potentials.

In the present chapter, we investigate how our results can be mapped to traditional network models with spiking dynamics. We specifically study whether the rate dynamics can be used to shed light on the phenomena which have been numerically reported in networks of leaky integrate-and-fire neurons, and cannot be explained in terms of classical mean field approaches [95].

As mentioned in Chapter 1, when the overall synaptic coupling is small, traditional mean field theories for spiking networks correctly predict an asynchronous dynamical regime, where neurons fire Poisson-like action potentials with stationary and homogeneous firing rate [24]. At high coupling strength, the equilibrium firing rate undergoes an instability to strongly fluctuating activity, which cannot be captured by the traditional mean field theory.

Classical mean field approaches for spiking networks differ from DMF as they provide a self-consistent description only at the level of the mean firing rate. The firing rate of the network can be determined as the fixed point of a self-consistent equation which includes the biophysical details of the network model [9]. Around those firing rate equilibrium points, rate dynamics has been proved to provide a crude but sometimes effective approximation of spiking activity [96, 114, 95]. In the approximated description, the spiking network is replaced by a network of rate units with exactly the same excitatory-inhibitory architecture.

In the same spirit, here we show that a simple network of rate units, whose input-to-rate activation function has been design to capture the main mechanisms of spike initiation, is able of reproducing the main qualitative features that have been observed in simulations of spiking network models. We interpret the results in terms of the two different regimes of fluctuating activity that have been found in Chapter 3 and we predict that, at high coupling strength, the average network firing rate is mostly controlled by the value of the refractory period. We confirm this prediction by performing direct numerical simulations in networks of leaky integrate-and-fire units.

4.1 Rate networks with a LIF transfer function

We focus again on the fixed in-degree synaptic matrix in which the inputs to excitatory and inhibitory neurons are statistically equivalent, but consider a rate network in which the dynamics are now given by:

$$\dot{\phi}_i(t) = -\phi_i(t) + F(\mu_i(t), \sigma_i(t)) \tag{4.1}$$

where:

$$\begin{aligned} \mu_i(t) &= \mu_0 + \tau_m \sum_j J_{ij}\phi_j(t) \\ \sigma_i^2(t) &= \tau_m \sum_j J_{ij}^2 \phi_j(t) \end{aligned} \tag{4.2}$$

Here ϕ_i is the firing rate of unit i, μ_0 is a constant external input, and $\tau_m = 20$ ms is the membrane time constant. The function $F(\mu, \sigma)$ is the input-output function of a leaky integrate-and-fire neuron receiving a white-noise input of mean μ and variance σ [125, 101]:

$$F(\mu, \sigma^2) = \left[\tau_{rp} + 2\tau_m \int_{\frac{V_r-\mu}{\sigma}}^{\frac{V_{th}-\mu}{\sigma}} \mathrm{d}\, u e^{u^2} \int_{-\infty}^{u} \mathrm{d}\, \nu e^{-\nu^2} \right]^{-1} \tag{4.3}$$

where V_{th} and V_r are the threshold and reset potentials of the LIF neurons, and τ_{rp} is the refractory period.

The firing rate model defined in Eq. 4.1 is directly related to the mean field theory for networks of LIF neurons interacting through instantaneous synapses [25, 24, 95]. More specifically, the fixed point of the dynamics defined in Eq. 4.1 is identical to the equilibrium firing rate in the classical asynchronous state of a network of LIF neurons with an identical connectivity as the rate model [25, 24]. Eq. 4.1 can then be seen as simplified dynamics around this equilibrium point [96, 114]. A linear stability analysis of the fixed point for the rate model predicts an instability analogous to the one found in threshold-linear rate models. A comparison with a network of LIF neurons shows that this instability predicts a change in the dynamics in the corresponding spiking network, although there may be quantitative deviations in the precise location of the instability.

The dynamics of Eq. 4.1 have been analytically investigated only up to the instability [95]. To investigate the dynamics above the instability, we set $x_i(t) = \sum_{j=1}^{N} J_{ij}\phi_j(t)$, and rewrite the dynamics in the more familiar form:

$$\dot{x_i}(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij} F(\tau_m x_j(t), \sigma_j(t)) \tag{4.4}$$

The main novelty with respect to previously studied rate models is that the input-output transfer function F depends on the standard deviation σ_j of the input current to the unit j. A dependence on a time-varying σ_j is however difficult to include in the dynamical mean field approach. As a step forward, we fix σ_j to its average value independent of j and time, which corresponds to substituting all the firing rates with a constant effective value $\bar{\phi}$:

$$\sigma^2 \sim \tau_m \sum_j J_{ij}^2 \bar{\phi} = \tau_m J^2 (C_E + g^2 C_I)\bar{\phi} \tag{4.5}$$

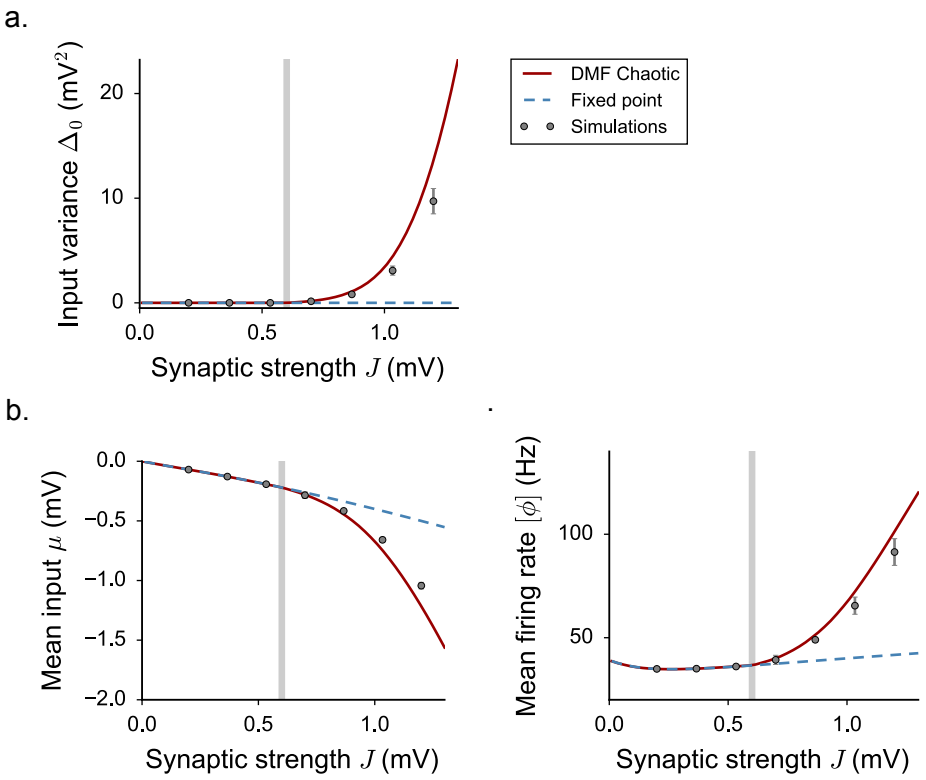


FIGURE 4.1: Dynamical mean field characterization of rate network activity with a LIF activation function, where we set $\sigma^2 = \tau_m J^2(C_E + g^2 C_I)\bar{\phi}$, $\bar{\phi} = 20$ Hz. **a-b-c.** Statistical characterization for $\tau_r = 0.5$ ms: input variance, mean input current and mean firing rate. Grey vertical lines indicate the position of the critical coupling. Choice of the parameters: $g = 5$, $C = 100$, $\mu_0 = 24$ mV.

With this substitution, we are back to a classical rate model with an LIF transfer function. Quantitatively the dynamics of that model are not identical to the model defined in Eq. 4.1, but they can be studied using dynamical mean field theory. We therefore focus on qualitative features of the dynamics rather than quantitative comparisons between models.

Solving the dynamical mean field equations shows that the dynamics in the rate model with and LIF transfer function are qualitatively similar to the threshold-linear rate model studied above. As the coupling strength J is increased above a critical value, the fixed point loses stability, and a fluctuating regime emerges. The amplitude of the fluctuations increases with coupling (Fig. 4.1 **a**), and induces an increase of the mean firing rate with respect to values predicted for the fixed point (Fig. 4.1 **c**).

In the LIF transfer function, the upper bound on the firing rate is given by the inverse of the refractory period. For that transfer function, changing the refractory period does not modify only the upper bound, but instead affects the full function. For different values of the refractory periods, the fixed point firing rate and the location of the instability therefore change, but these effects are very small for refractory periods below one millisecond.

Varying the refractory period reveals two different fluctuating regimes as found in threshold-linear rate models (Fig. 4.2 **a-b-c**). At intermediate couplings, the fluctuating dynamics de-

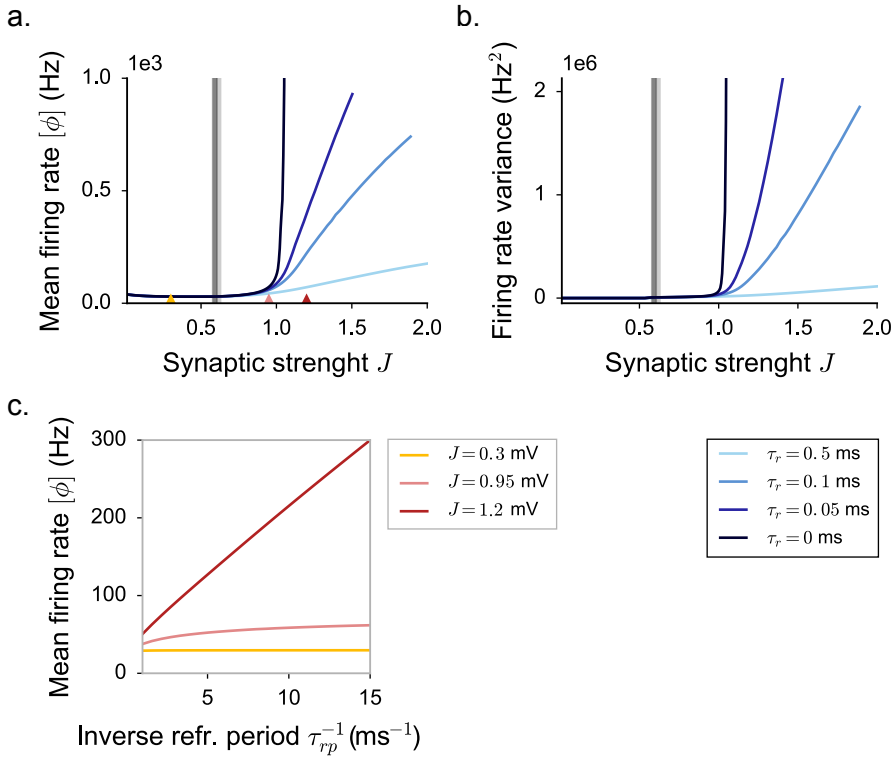


Figure 4.2: Dynamical mean field characterization of rate network activity with a LIF activation function, appearance of the two regimes. **a-b.** Mean firing rate and rate standard deviation for different values of the refractory period, determining slightly different positions of the transition (grey lines). Choice of the parameters: $g = 5$, $C = 100$, $\mu_0 = 24$ mV. **c.** Mean firing rate dependence on the refractory period, the inverse of which determines the saturation value of the transfer function. The three values of the synaptic coupling, indicated by triangles in **a**, correspond to the three different regimes.

pend weakly on the refractory period and remain bounded if the refractory period is set to zero. At strong couplings, the fluctuating dynamics are stabilized only by the presence of the upper bound, and diverge if the refractory period is set to zero. The main difference with the threshold-linear model is that the additional dependence on the coupling J induced by σ on the transfer function reduces the extent of the intermediate regime.

4.2 Spiking networks of leaky integrate-and-fire neurons: numerical results

Having established the existence of two different regimes of fluctuating activity in rate networks with an LIF transfer function, we next consider spiking networks of LIF neurons. To compare the different regimes of activity in spiking networks with the regimes we found in rate networks, we performed direct numerical simulations of a spiking LIF network.

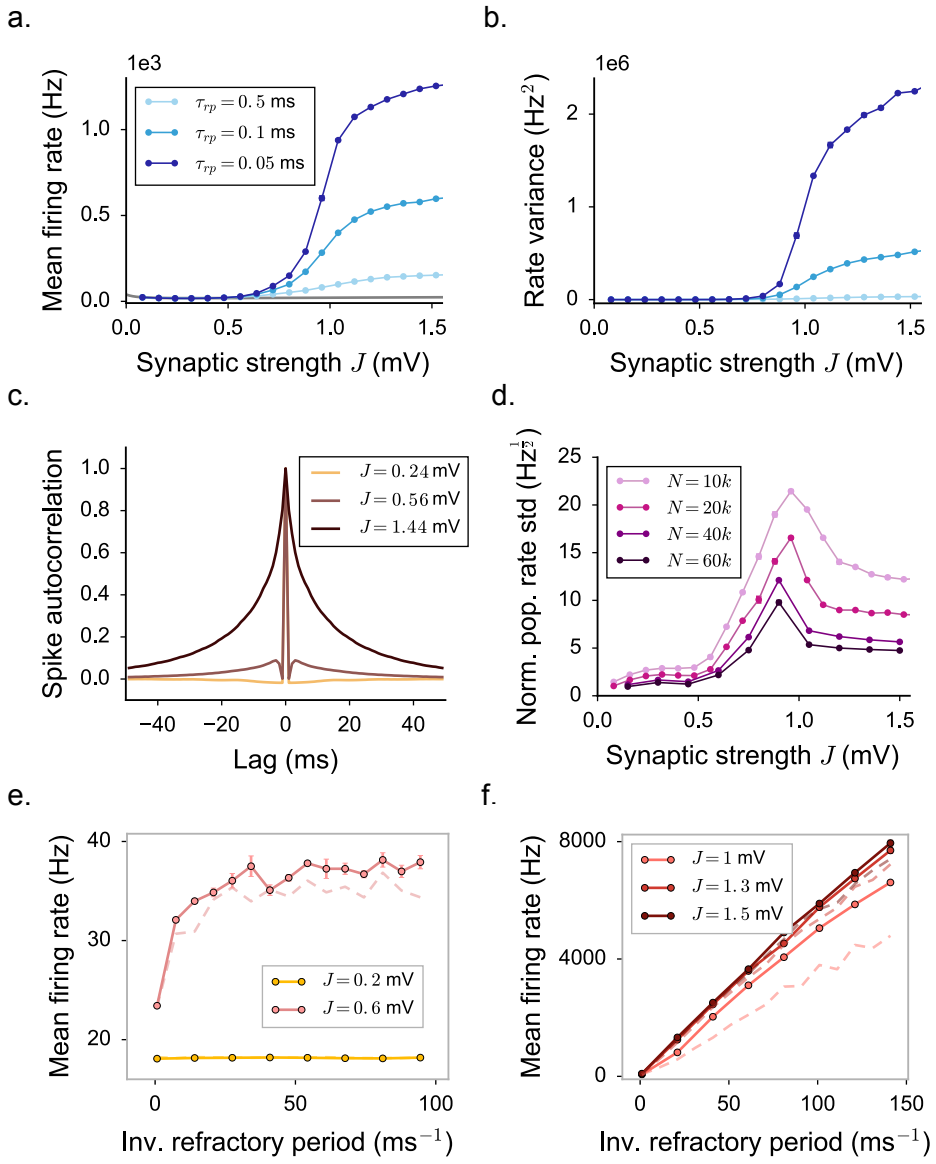


FIGURE 4.3: Statistical characterization of activity in a network of leaky integrate-and-fire neurons. **a.** Mean firing rate. Numerical simulations ($N = 20000$) are in good agreement with the LIF mean field prediction (grey line) for low coupling values ($J < 0.5$). For high values of J ($J > 0.8$), mean firing rates diverge and becomes highly dependent on the refractory period. **b.** Firing rate variance, computed on instantaneous firing rates evaluated with a 50 ms Gaussian filter. **c.** Spike autocorrelation function, computed with 1 ms time bins, for three different values of the coupling J ($\tau_{rp} = 0.5$). **d.** Dependence on J and N of correlations and synchrony, quantified by the std of the population-averaged spiking rate, normalized by the square root of the mean firing rate ($\tau_{rp} = 0.05$). Std is computed within a time bin of 1 ms. **e-f.** Direct dependence between the mean firing rate and refractory period. Panel **e** shows the low and intermediate coupling regime. Panel **f** shows the high coupling regime. Colored dots: simulated networks with $N = 20000$. Lighter dashed lines (when visible) show the result for $N = 10000$. In all the panels, choice of the parameters: $g = 5$, $C = 500$, Δ=1.1 ms, $\mu_0 = 24$ mV.

The membrane potential dynamics of the i-th LIF neuron are given by:

$$\tau_m \frac{dV_i}{dt} = -V_i + \mu_0 + RI_i(t) + \mu_{\text{ext}}(t) \tag{4.6}$$

where $\tau_m = 20$ ms is the membrane time constant, μ_0 is a constant offset current, and RI_i is the total synaptic input from within the network. When the membrane potential crosses the threshold $V_{\text{th}} = 20$ mV, an action potential is emitted and the membrane potential is reset to the value $V_{\text{r}} = 10$ mV. The dynamics resume after a refractory period τ_{r}, the value of which was systematically varied. The total synaptic input to the i-th neuron is:

$$RI_i(t) = \tau_m \sum_j J_{ij} \sum_k \delta(t - t_j^{(k)} - \Delta) \tag{4.7}$$

where J_{ij} is the amplitude of the post-synaptic potential evoked in neuron i by an action potential occurring in neuron j, and Δ is the synaptic delay (here taken to be 1.1 ms). Note that if the synaptic delay is shorter than the refractory period, the network develops spurious synchronization. The connectivity matrix J_{ij} was identical to the rate network with fixed in-degree described above.

We examined the effects of the coupling strength and refractory period on first- and second-order statistics (Fig. 4.3 **a-b**), i.e. the mean firing rate and the variance of the activity (computed on instantaneous firing rates evaluated with a 50 ms Gaussian filter).

For low couplings strengths, the mean firing rate in the network is close to the value predicted for the fixed point of Eq. 4.1, i.e. the equilibrium asynchronous state, and essentially independent of the refractory period. Similarly, the variance of the activity remains at low values independent of the refractory period. As the synaptic strength is increased, the mean firing rate deviates positively from the equilibrium value (Fig. 4.3 **a**), and the variance of the activity increases (Fig. 4.3 **b**). For intermediate and strong synaptic coupling, the values of first- and second-order activity statistics become dependent on the values of the refractory period.

Specifically, for intermediate values of the coupling, the mean firing rate increases with decreasing refractory period, but saturates with decreasing refractory period (Fig. 4.3 **e**). This is similar to the behavior of the rate networks in the inhibition-stabilized fluctuating regime. For large values of the coupling, the mean firing rate instead diverges linearly with the inverse of the refractory period (Fig. 4.3 **f**), a behavior analogous to rate networks in the second fluctuating regime in which the dynamics are only stabilized by the upper bound on the activity. The strength of the sensitivity to the refractory period depends on the inhibitory coupling: the stronger the relative inhibitory coupling, the weaker the sensitivity to the refractory period.

The main qualitative signatures of the two fluctuating regimes found in networks of rate units are therefore also observed in networks of spiking LIF neurons. It should be however noted that the details of the dynamics are different in rate and LIF networks. In particular, the shape of auto-correlation functions is different, as LIF neurons display a richer temporal structure at low and intermediate coupling strengths. At strong coupling, the auto-correlation function resembles those of rate networks with spiking interactions (see Fig. 3.5 **c**), in particular it displays a characteristic cusp at zero time-lag. The simulated LIF networks show no sign of critical slowing down, as expected from the analysis of the effects of spiking noise on the activity.

Moreover, strong finite-size effects are present in the simulations. To quantify correlations among units and synchrony effects deriving from finite-size effects, we measure the standard

deviation of the amplitude of fluctuations in the population-averaged activity, normalized by the square root of the mean firing rate (Fig. 4.3 **d**). Correlations and synchrony appear to be stronger for small values of the refractory period. The effect of correlations is furthermore weaker in the low and high coupling regimes, and it has a maximum for intermediate couplings. However, whatever the value of J, they decay as the system size is increased (for a more detailed characterization, see Appendix B).

In summary, for the range of values of the refractory period considered here, the activity in a network of spiking neurons is in qualitative agreement with predictions of the simple rate models analyzed in the previous sections. The rate model introduced in Eq. 4.1 however does not provide exact quantitative predictions for the firing rate statistics above the instability. In particular, due to the numerical limitations in considering the limit $\tau_{rp} \to 0$, it is not possible to evaluate exactly through simulations the position of an equivalent critical value J_D.

4.3 Discussion

How a regime analogous to rate chaos appears in networks of integrate-and-fire neurons has been a topic of intense debate. Two different scenarios have been proposed: (i) rate chaos appears in networks of spiking neurons only in the limit of very slow synaptic or membrane time-constants [58]; (ii) rate chaos appears in generic excitatory-inhibitory networks, i.e. for arbitrarily fast synaptic time-constants [95]. The heart of the debate has been the nature of the signature of rate chaos.

The classical signature of the transition to rate chaos is *critical slowing-down*, i.e. the divergence of the timescale of rate fluctuations close to the critical coupling [127]. As it has been shown (see also parallel studies [55, 69]), spiking interactions induce noise in the dynamics, and critical slowing down is very sensitive to the amplitude of such noise. The amplitude of this spiking noise is moreover proportional to $1/\sqrt{\bar{\tau}}$, where $\bar{\tau}$ is the timescale of the rate model, usually interpreted as the slowest timescale in the system (either membrane or synaptic timescale). Critical-slowing down can therefore be observed only when the membrane or synaptic timescales are very slow and filter out the spiking noise [58, 69].

Here we have shown that for networks with E-I connectivity and positive firing rates, a novel signature of fluctuating activity appears simply at the level of mean and variance of firing-rates, which become highly sensitive to the upper bound at strong coupling. In contrast to critical slowing-down, this signature of strongly fluctuating activity appears to be very robust to noise, and therefore independent of the timescale of the synapses or membrane time constant. Simulations of networks of integrate-and-fire neurons reveal such signatures of underlying fluctuating activity for arbitrarily fast synaptic time-constants, although no critical slowing down is seen or expected.

The results presented here therefore reconcile the two proposed scenarios: a sharp phase-transition to fluctuating activity characterized by critical slowing down appears only in the limit of very slow synaptic or membrane time-constants; a smooth cross-over to strongly fluctuating activity can however be observed for arbitrarily fast synaptic time-constants.

4.3.1 Mean field theories and rate-based descriptions of integrate-and-fire networks

The dynamical mean field theory used here to analyze rate networks should be contrasted with mean field theories developed for integrate-and-fire networks. Classical mean field theories

for networks of integrate-and-fire neurons lead to a self-consistent firing rate description of the equilibrium asynchronous state [9, 25, 24], but this effective description is however not consistent at the level of the second order statistics. Mean field theories for IF neurons assume indeed that the input to each neuron consists of white noise, originating from Poisson spiking; however the firing of an integrate-and-fire neuron in response to white-noise inputs is in general not Poisson [94], so that the Poisson assumption is not self-consistent. In spite of this, mean field theory predicts well the first-order statistics over a large parameter range [57], but fails at strong coupling when the activity is strongly non-Poisson [95].

Extending mean field theory to determine analytically self-consistent second-order statistics is challenging for spiking networks. Several numerical approaches have been developed [77, 47, 149], but their range of convergence appears to be limited. A recent analysis of that type has suggested the existence of an instability driven by second-order statistics as the coupling is increased [149].

A simpler route to incorporate non-trivial second order statics in the mean field description is to describe the different neurons as Poisson processes with rates that vary in time. One way to do this is to replace every neuron by a linear-nonlinear (LN) unit that transforms its inputs into an output firing rate, and previous works have shown that such an approximation can lead to remarkably accurate results [96, 137, 97, 114]. If one moreover approximates the linear filter in the LN unit by an exponential, this approach results in a mapping from a network of integrate-and-fire neurons to a network of rate units with identical connectivity [95]. Note that such an approximation is not quantitatively accurate for the leaky integrate-and-fire model with fast synaptic timescales - indeed the linear response of that model contains a very fast component ($1/\sqrt{t}$ divergence in the impulse response at short times, see [96]). A single timescale exponential however describes much better dynamics of other models, such as the exponential integrate-and-fire [96]. The accuracy of the mapping from integrate-and-fire to rate networks also depends on synaptic timescales which influence both the amplitude of synaptic noise and the transfer function itself [26]. It has been argued that the mapping becomes exact in the limit of infinitely long timescales [123, 58].

In this study, we have analyzed rate networks using dynamical mean field theory. This version of mean field theory is different from the one used for integrate-and-fire networks as it determines self-consistently and analytically not only the first-order statistics, but also the second-order statistics, i.e. the full auto-correlation function of neural activity. Note that this is similar in spirit to the approach developed for integrate-and-fire networks [77, 47, 149], except that integrate-and-fire neurons are replaced by simpler, analytically tractable rate units. Dynamical mean field theory reveals that at large coupling, network feedback strongly amplifies the fluctuations in the activity, which in turn lead to an increase in mean firing rates, as seen in networks of spiking neurons [95]. The rate-model moreover correctly predicts that for strong coupling, the activity is highly sensitive to the upper bound set by the refractory period, although the mean activity is well below saturation.

As pointed out above, the mapping from an integrate-and-fire to a rate network is based on a number of approximations and simplifications. The fluctuating state in the rate network therefore does not in general lead to a quantitatively correct description of the activity in a network of integrate-and-fire neurons. However, the rate model does capture the existence of a fundamental instability, which amplifies fluctuations through network feedback.

Part II

RANDOM NETWORKS AS RESERVOIRS

Computing with recurrent networks: an overview 5

Recurrent networks of intricately connected cells represent the elementary units of computation in the cortex. Computations can be thought as specific – possibly non-linear – input-output relationships which are implemented at the population level by orchestrating together a wide spectrum of single cells responses.

The dynamical mechanisms which support computations in cortical recurrent networks can be investigated by building artificial network models which are able to satisfy the input-output rules specified by different tasks. Historically, computational network models have been constructed by exploiting two main approaches. The network structure can be explicitely designed by following a simple, and often low-dimensional, theoretical intuition; in alternative, it can be obtained algorithmically by training unspecialized network models on synthetic data [13]. Traditional hand-crafted models are often easier to analyse, but suffer from the drawback of being hard to match with the high degree of spatial and temporal disorder that has been observed from in-vivo recordings of cortical activity. On the other hand, computations and complexity appear to be naturally combined in the network models which come out of algorithmic training procedures, but the exact dynamical mechanisms on which they are built are often hard to capture.

In this chapter, we review the main ideas, results and limitations of both approaches. In the spirit of providing a quantitative understanding of the dynamics within recurrent computational networks, the two frameworks are adopted and mixed together in the two parts which constitute the rest of this dissertation.

5.1 Designing structured recurrent networks

In the theoretical literature related to the study of dynamics in recurrent neural networks, random networks have historically gained a prominent position [127, 24, 143, 105]. In a random network, there are no preferred units with specified input or functional role: the activity profile of every neuron in the population can be thought instead as randomly extracted from a continuous probability distribution. For this reason, the neural activity can be characterized with purely statistical approaches, which considerably simplify the analysis [127, 24].

As discussed in Chapter 1, random networks have become a fruitful theoretical paradigm

that has led to the development of fundamental concepts such as excitation-inhibition balance [24, 143] and decorrelation [105]. However, randomly connected recurrent networks display only very stereotyped responses to external inputs and can implement only a limited range of input-output computations [99, 74]. Moreover, the connectivity structure of real cortical circuits is thought to possess significantly non-random features [59, 129].

Networks of excitatory and inhibitory neurons constitute the fundamental computational units of the cortex. A central hypothesis in neurobiology states that cortical computations, in the form of arbitrary input-output associations, emerge from precise patterns in the synaptic connectivity. Connectivity structures are created and continuously reshaped, at different time scales, from a variety of synaptic plasticity mechanisms [38, 1].

Directly measuring the strength of the synaptic connections among pairs of neurons is a complex experimental process which can give access to a limited fraction of information [129, 71]. As a consequence, the strategies that the brain adopts to solve and implement its tasks are mostly unknown. To fill this gap, synthetic network models, which are able to solve some among the tasks that cortical networks are likely to face, can be built and analyzed.

Traditionally, ad-hoc network models have been explicitly designed by combining intuition with experimental insights. This class of hand-crafted models consists of connectivity structures that, associated to simple single neuron models, returns circuits which correctly implements the desired task. This approach results in task-specific models with little degree of flexibility.

In hand-crafted computational models, the strengths of all synapses and the activity of all neurons are known, yet an understanding of the relation between connectivity, dynamics and computations has been achieved only in very specific cases. Hand-crafted recurrent networks are typically large-size implementations of discrete [65, 148] or continuous [118, 17, 29, 124, 81] dynamical attractors, that have been constrained to reproduce some distinctive features of the recorded neural activity. One well-known example is given by the network models for working memory [109, 50], where a subset of similarly tuned units display sustained activity in absence of driving inputs [147, 152, 81].

Some effort has been devoted to reconciling hand-crafted networks with the large neural variability which has been observed in data. In some cases, the original connectivity structure – often simple and homogeneous – has been combined with disordered synapses, resulting in network models where computations co-exists with Poisson-like firing [147, 110, 104, 121]. It is more difficult, however, to generalize hand-crafted architectures to other more recent experimental findings.

In most of hand-crafted network models, neurons are identified by their tuning properties. Units with similar tuning have highly homogeneous average activity profiles. On the contrary, recent advances in the recording techniques have revealed that neural responses are much more complex and heterogeneous than what hypothesized decades ago [32, 107]. Although a minor fraction of cells display clear and stereotyped tuning properties, the response profile of the majority of neurons is typically found to be mixed, non-stationary and multi-phasic (Fig. 5.1) [32, 23, 83, 37, 107]. Broad and non-stationary selectivity properties might contribute to enrich the computational capacity of standard network models, as they project the neural activity in a higher-dimensional space where simple computations like discrimination can be more robustly performed [107, 14].

Heterogeneous and non-stationary responses cannot be trivially incorporated in the highly simplified computational structure of hand-crafted networks. Very recently, a complementary line of research, which aims at building computational models by algorithmic approaches

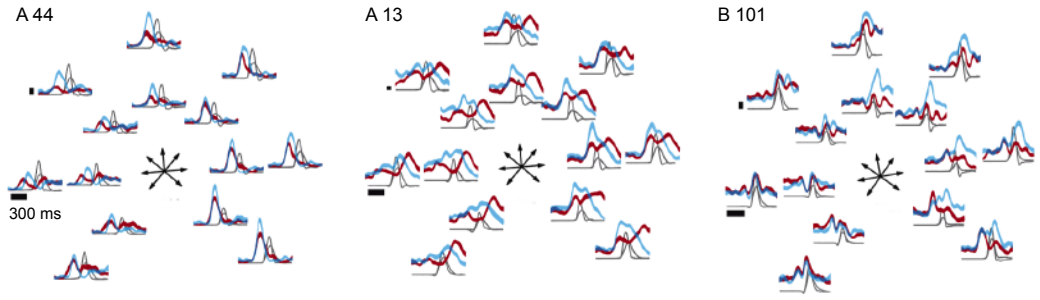


FIGURE 5.1: Tuning properties of three sample neurons from the motor cortex of two monkeys A and B, while the animals are performing a reaching task. Seven reaching directions have been tested. Internal and external panels correspond to two different fixation-target distances. Purple and blue traces display the inter-trial average firing rate for fast and slow reach movements. Grey traces plot the mean hand velocity, which is comparable by experimental design. Vertical calibration bars indicate 20 spikes/s. Adapted from [32].

rather than by inspection, has been proposed and successfully tested in a variety of different scenarios [15, 13].

5.2 Training structured recurrent networks

Computational network models are circuit implementations of specific tasks. Once the task has been modeled as a specified input-output transformation, random neural networks can be trained on synthetic data by trial and error. The training algorithm, which is iterated up to convergence, determines how synaptic modifications should be computed in every trial.

Because of the intricate temporal dependencies generated by recurrent dynamics, designing efficient learning algorithms for recurrent neural networks has ben historically a cumbersome task. Indeed, the traditional approaches that have permitted, in the eighties, to efficiently train multi-layered feedforward architecture [112] do not readily generalize to recurrent models [18]. More refined and alternative solutions have been developed in the last fifteen years, leading to a new fruitful research line at the intersection between machine learning and computational neuroscience. In the following, we introduce in detail the computing architectures which will be considered in the rest of Parts II and III.

5.2.1 Reservoir computing

A clever strategy to circumvent the problem of dealing with long-standing temporal feedbacks was proposed sixteen years ago in the parallel works of Jaeger [67] and Maass [79].

Both approaches aim at exploiting to the maximal degree the rich dynamics which is intrinsically generated in large random architectures. To this aim, only a minimal set of synaptic connections is exposed to plasticity. Spontaneous activity acts as a reservoir of basis function from which a wide set of target responses can be reconstructed. For this reason, both methods commonly undergo under the name of *reservoir computing*.

In the following, we focus on the continuous-time formulation of the network setup by [67, 68]. A random network of rate units, equivalent to the model in [127] (see also Chapter 2), is trained to associate a given time-varying output $f(t)$ to a given input function $I(t)$ (Fig. 5.2

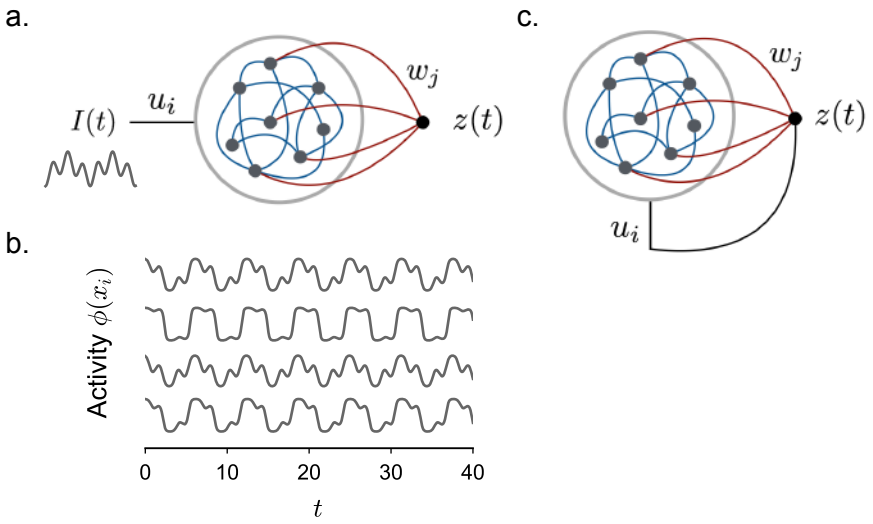


Figure 5.2: Reservoir training: architecture and principles of echo-state machines [67]. **a.** Original open-loop reservoir architecture. A time-varying input current is injected in a random network, and the input-evoked response is used as a set of basis functions to reconstruct the target function $f(t)$. The output is defined at the level of a linear readout of signal $z(t)$. **b.** The input-driven response, time traces of four arbitrarily selected units. The input current consist of a sum of two sinusoidal functions. The random network response is more temporally complex, and strongly heterogeneous. This response can be used to approximate a huge number of different periodic functions. **c.** Reservoir closed-loop architecture: the linear output $z(t)$ is used itself as a forcing input.

a). The reservoir dynamics read:

$$\dot{x_i}(t) = -x_i(t) + g\sum_{j=1}^{N} \chi_{ij}\phi(x_j(t)) + u_i I(t), \tag{5.1}$$

where the input weights u_i are $\mathcal{O}(1)$ and can be set at random. The network output is defined at the level of a readout signal $z(t)$, where the network activity is linearly combined using a decoding vector w:

$$z(t) = \sum_{j=1}^{N} w_j \phi(x_j(t)). \tag{5.2}$$

In contrast to traditional hand-crafted models, computations in reservoir machines are based on the input-dependent time course of neural trajectories, which involve both transient and attractor dynamics [28].

Solely the outgoing weights w_j are tuned during learning, which makes training a particularly easy task. From the point of view of the output unit, the architecture presents in fact no feedback loops. Since the input function is fixed, the input-driven response of the random network can be determined by simply integrating the dynamics over time. Because of the disordered and non-linear dynamics of the reservoir, the activity profiles of different units reflect the input dynamics but display more heterogeneous and rich temporal properties (Fig. 5.2 **b**). Once the time traces of neural activity have been stored, Eq. 5.2 can be thought as a set of linear constraints - one for every point in time - on the outgoing weights w_j. If the input and the

output function are well matched, the system of equations admits at least one solution, which can be computed by inverting Eq. 5.2 through standard numerical techniques like least-square minimization [67].

In order to avoid unpredictable network responses, traditional reservoirs operate out of the chaotic regime. As discussed in Chapter 2, in absence of external input currents, this condition is equivalent to imposing overall connectivity strengths g smaller than unity. As observed in a variety of different frameworks, however, external inputs have the effect of further stabilizing stationary and time-varying attractors [99, 89, 86, 87], so that in practice milder conditions on g need to be imposed. As the training procedure directly exploits the spontaneous input-driven dynamics of the network, no further subtle stability issues must be considered.

The domain of target functions which can be reconstructed at the output level is tightly constrained by the temporal structure of the input. Nevertheless, the recurrent circuitry of the reservoir can be used to significantly slow down the transient reservoir response. As shown in Chapter 2, the largest decay time scale of a random reservoir increases with g, and diverges at the boundary with the chaotic regime. It has been proposed that reservoir machines which have been tuned right below the instability to chaos possess slow response properties which turn into optimal temporal computational capacities [19, 76].

5.2.2 Closing the loop

For certain input-output pairs, appropriately tuning the reservoir parameters is not sufficient to circumvent the problem. Consider for example the case of a pattern generator network which associates to a stationary input level an output sinusoid with the same frequency. In cases like that one, additional inputs which possess the spatial-temporal structure of the outputs must be considered.

One parsimonious option consists of exploiting the output signal itself as forcing input function [67, 80]. When the output is used as input, a feedback element is introduced in the architecture (Fig. 5.2 **c**). The reservoir dynamics transforms into:

$$\dot{x_i}(t) = -x_i(t) + g\sum_{j=1}^{N}\chi_{ij}\phi(x_j(t)) + u_i z(t), \tag{5.3}$$

which can be combined with Eq. 5.2 to give:

$$\dot{x_i}(t) = -x_i(t) + \sum_{j=1}^{N}(g\chi_{ij} + u_i w_j)\phi(x_j(t)). \tag{5.4}$$

The global feedback architecture can thus be interpreted as a novel recurrent architecture, where the random connectivity has been summed to a one-dimensional structured element.

Because of the feedback, any change in the output weights w_j results in finite perturbations to the reservoir dynamics. As a consequence, the training procedure which applies to open-loop architectures does not directly extend to closed-loop networks.

In order to find a proper training strategy we observe that, in the final trained network, the unique feedback signal $z(t)$ must faithfully reproduce the target $f(t)$. Once the feedback has been clamped by setting $z(t) = f(t)$, the reservoir response is univocally determined. The output weights w_j can thus be determined as in the open-loop case by properly combining the reservoir activity profiles into the output target function.

Even if the training procedure remains simple enough, introducing feedback readout units comes at the price of generating more complex, and hard to predict, recurrent dynamics. The batch training procedure returns in fact a network model which admits a self-consistent solution where the readout signal $z(t)$ coincide with the target. It does not allow to control, however, the number and the stability properties of all the dynamical attractors that are generated with the weights update. As a consequence, the spontaneous dynamics of the final network model might deviate from the target because of instability or multi-stability issues.

Training can be improved by transforming batch into *online* update rules. Online procedures [132, 73, 45] force the synaptic updates to take into account the spontaneous recurrent dynamics of the closed-loop network. Progressive weights modifications occur indeed together with the integration in time of the network activity. As a result, some among the dynamical instabilities can be sampled and avoided. In contrast to batch techniques, during online training the learning algorithm is iterated over a large number of steps, until synaptic modifications reach convergence.

Online training algorithms have been applied to different network architectures. In its original formulation, the FORCE training algorithm [132] has been used mostly on feedback reservoir architectures similar to [67]. In most of the later training procedures, plasticity has been extended instead to a larger subsample or to the totality of the N^2 synapses of the recurrent network [73, 45].

5.2.3 Understanding trained networks

Together with novel optimization techniques [85], online training algorithms have been successfully adopted in the last few years to construct a new class of structured networks models.

Trained neural networks implements neural computations by means of heterogeneous and non-stationary activity patterns [15, 134], which can in some cases provide a better explanation of neural data than traditionally designed architectures. The algorithmic approach is furthermore beneficial when dealing with complex behavioural tasks, for which hand-crafted implementations are very difficult to design [83]. Finally, random networks can be trained on in-vivo neural recordings, which opens the possibility to directly infer the synaptic connectivity schemes that the brain adopts to implement specific tasks [100].

In contrast to the simple and homogeneous synaptic schemes from hand-crafted models, the connectivity structures emerging from training are extremely complex and hard to interpret. The general computational principles which underly the task implementation are furthermore hard to isolate, as the network units often display mixed and time-varying tuning properties.

From a more theoretical perspective, the dynamical principles underlying asymptotic and transient activity in trained networks are most of the time obscure as well [13]. Very recently, progress has been made in reverse-engeneering, both numerically [133] and theoretically [108], the specific network architectures which directly emerge from training. Apart from reverse-engeneering, transforming trained recurrent networks into a formalized computational framework requires to answer more general theoretical questions lying at the intersection between the fields of network dynamics, random matrix and control theory. A unified theoretical setup would allow to answer fundamentally unsolved questions like: are there limitations in the computational power of the different network architectures? Which are the input-output relationships that their dynamics can implement? Which are the optimal parameters which allow good training performances?

In the spirit of developing a theoretical understanding of trained networks, reservoir ar-

chitectures might represent a particularly fertile field of investigation [108]. As most of the synaptic connectivity is random, reservoir networks might partly benefit from the standard statistical approaches which have been developed for purely random networks.

Already at the level of random network dynamics, a series of open problems still waits to be exactly quantified. For example, although the interactions between external inputs and chaotic activity have been addressed in many simplified scenarios [89, 99, 69, 87], very little is known about the general properties of time-varying responses in setups which more directly apply to reservoir machines [86]. A step in this direction is taken in the next chapter of this thesis.

Summary of Chapter 6

Neural activity in strongly connected random networks is highly heterogeneous and dynamically complex. One recent line of research suggests that, perhaps counter-intuitively, such disordered states are highly desirable if the network aims at performing robust and high-precision spatio-temporal tasks. If the dynamics are driven in a sub-instability regime where chaotic fluctuations are suppressed by the external inputs, activity traces can be used as a rich basis set from which a large number of complex target functions can be reconstructed. This idea has been developing as a new machine learning set of techniques under the name of reservoir computing.

Very few studies have focused on the dynamical mechanisms which allow such training procedures to be effective. The result is a very poor understanding of successes and failures which are obtained when training with those algorithms. In order to start developing a more rigorous understanding of those computational frameworks, we perform a quantitative analysis of a feedback reservoir architecture which is trained to reproduce periodic output functions.

In analysing the existence and the degeneracy of the global solutions, we find that a critical role is played by the degree of synchronization with which recurrent random networks respond to synchronized inputs. We show that, within a simple linear network, weak synchronization can be achieved with low frequency target functions and highly disordered connectivities, which set the network dynamics just below the boundary to instability. Remarkably, common training algorithms like the recursive least-square minimization seem to efficiently take advantage from desynchronized reservoir activity. Numerical investigations, furthermore, suggest that weakly-coupled non-linear networks display response properties and performances which are comparable with the simplified linear models. Our analysis thus provide a good description of trained non-linear networks in the low-coupling strength regime.

The main results of this part of the dissertation derive from a work which has been conducted as a Summer School project, held in Woods Hole in 2015 (MCN: Methods in Computational Neuroscience). The project has been done under the supervision of H. Sompolinsky and R. Rubin.

Analysis of a linear trained network 6

A quantitative analysis of trained recurrent networks is a complex problem which can be approached in several different ways. From this perspecive, some simple recurrent setups, like the feedback architecture which is widely adopted in [68, 132], might represent a particularly fertile ground.

One possibility is to zoom in, after learning, on the portion of the neural circuit contained inside the reservoir. As its recurrent connectivity is not affected by synaptic plasticity mechanisms, from a technical point of view such a network is purely random. As a consequence, mean field approaches could indicate a way to characterize quantitatively its response properties and the expected dynamical regimes.

In this chapter, we perform a quantitative analysis of a feedback architecture which is trained to reconstruct periodic output functions. We show that understanding how random reservoir networks respond to synchronized inputs is a crucial step in defining the limitations and the computational properties of those architectures.

We look for exact and statistical characterizations of network activity in the case of a linear input-to-rate activation function. We show that already such a simple setup presents nontrivial resonances and synchronization properties, which suggest a an optimal range of parameters for training algorithms performances.

6.1 From feedback architectures to auto-encoders and viceversa

We focus on a single feedback reservoir architecture, inspired from [68, 132] (see also Chapter 5). The output signal $z(t)$ is linearly extracted from a random reservoir of rate units, and the same signal is provided as a feedback to the neural population through some arbitrary weights u_i (Fig. 6.1 **a**). The evolution of the reservoir activity follows the usual rate dynamics (Eqs. 2.1 and 5.3) [127], and we indicate with $J_{ij} = g\chi_{ij}$ its random connectivity. The decoding set is the only synaptic component which is affected by learning, and it is indicated with $\{w_j\}$.

During learning, the readout signal $z(t) = \sum_j w_j \phi(x_j(t))$ is trained to reproduce a certain target function $f(t)$. In this study, we consider a periodic function $f(t)$, and we expand it in a finite number of sinusoidal components.

We look for a rigorous description of the whole trained system which would help to answer

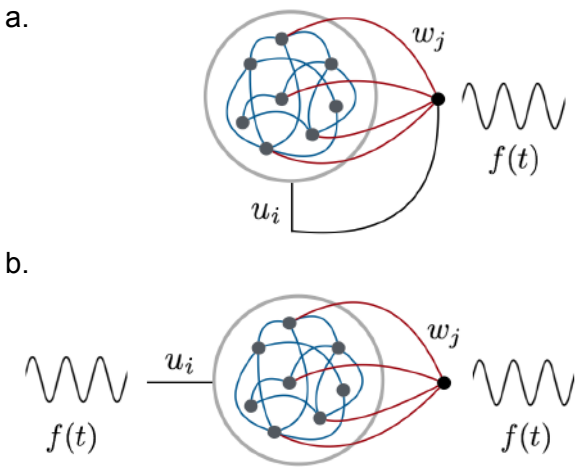


Figure 6.1: Network architectures considered in this analysis. **a.** Feedback architecture from [132]. The only learnt synaptic weights are the decoding ones, and they are drawn in red. After learning, the readout signal $z(t)$ coincides with the target function, which in the figure is a simple sinusoid. **b.** The corresponding auto-encoder architecture, where the feedback wiring is cut and the target function $f(t)$ acts both as input and as output.

the following questions: how is this configuration sustainable for the global network dynamics? Which kind of target functions are admitted, and which values of the network parameters need to be adopted?

To this end, we imagine to set our analysis at the end of the learning procedure, such that we can safely assume $z(t) = f(t)$. We then transform the feedback architecture into an open-loop, auto-encoder setup (Fig. 6.1 **b**). After training, the original reservoir is equivalent to a random network which receives the target function $f(t)$ as an external driving input. This input reshapes deterministically the reservoir activity into a temporal response $\{\phi(x_i(t))\}$. In a second step, activity is decoded to reconstruct back the target function, through $f(t) = \sum_j w_j \phi(x_j(t))$. The analysis is thus conducted in two steps: in the encoding phase, we determine the activity response $\{\phi(x_i(t))\}$; in the decoding phase, we use the solution to constrain the output weights and solve the decoder equation as a function of the N variables w_j. The auto-encoder formulation allows to assess the existence and compute the degeneracy of the solution $\{w_j\}^*$.

We observe that, both in linear and non-linear network architectures, any solution to the open-loop problem is not guaranteed to be stable with respect to the global dynamics of the closed feedback network (see, for example, [108] and Part III). However, as it will be shown in the following paragraphs, global stability is particularly easy to assess in the simplified network model that we are going to consider.

6.1.1 Exact solution

We start by solving the encoding step, which corresponds to study how a random network responds to synchronized temporal signals. The external input entering different units is given by $u_i f(t)$, and is thus highly correlated in time from one unit to the other. Several studies have investigated the dynamics of continuous-time random networks in presence of external

inputs, but they have focused on the case of decorrelated [99, 89, 55, 87] or stationary signals [69, 58, 87].

Synchronized inputs imply a non-trivial time course for the population-averaged activity [86]. This feature makes a quantitative analysis more complex from a technical point of view. We thus focus on the simpler case of linear rate models, and we set $\phi(x) = x$. With this choice, the solution to the encoding step can be computed exactly. Note that, for stability reasons (see Chapter 2), we need to restrict ourselves to weakly-coupled networks: $g < 1$.

We start from considering a purely sinusoidal target function: $f(t) = \cos(\omega_0 t)$. Network dynamics in the encoding step can be written as follows:

$$\dot{x}_i(t) = -x_i(t) + g \sum_{j=1}^{N} \chi_{ij} x_j(t) + u_i \cos(\omega_0 t). \tag{6.1}$$

Note that here, as in the previous chapters, we have rescaled time to set the time constant of integration to unity.

We decompose $x_i(t)$ on the basis given by the eigenvectors of the random matrix $g\chi_{ij}$. We indicate the eigenvectors set by $\{\vec{v}^l\}$, and the corresponding complex eigenvalues by $\{\lambda_l\}$. We obtain: $x_i = \sum_l x_l v_i^l$ and $u_i = \sum_l u_l v_i^l$. On this basis, Eq. 6.1 reads:

$$\dot{x}_l(t) = -(1 - \lambda_l) x_l(t) + u_l \cos(\omega_0 t) \tag{6.2}$$

whose asymptotic solution is given by:

$$x_l(t) = \frac{u_l}{2[(1-\lambda_l) + i\omega_0]} e^{i\omega_0 t} + \frac{u_l}{2[(1-\lambda_l) - i\omega_0]} e^{-i\omega_0 t}. \tag{6.3}$$

Notice that we expect both u_l and λ_l to be complex. Combining the eigenvectors in pairs of complex conjugates we can write the final expression for x_i as:

$$x_i(t) = \sum_{l'} X_{l'} \cos(\omega_0 t + \gamma_{l'}) \tag{6.4}$$

where the index l' runs over each pair of complex conjugates and each real eigenvalue. By summing over l', we conclude that the activity of single neurons is oscillating with frequency ω_0:

$$x_i(t) = A_i \cos(\omega_0 t + \varphi_i). \tag{6.5}$$

For every reservoir network χ_{ij}, the amplitudes A_i and the phases φ_i can be written in terms of the vector u_i and of the random eigenvalues and eigenvectors of the matrix χ_{ij}.

Once the deterministic network response has been computed, we turn to the decoder phase. We ask whether a proper set of weights $\{w_j\}$ exists such that we can write:

$$\begin{aligned} f(t) = \cos(\omega_0 t) &= \sum_{j=1}^{N} w_j A_j \cos(\omega_0 t + \varphi_j) \\ &= \sum_{j=1}^{N} w_j \left[A_j \cos(\omega_0 t) \cos(\varphi_j) - A_j \sin(\omega_0 t) \sin(\varphi_j) \right]. \end{aligned} \tag{6.6}$$

This equation can be interpreted in the complex plane as a system of two constraints (one for the real, and one for the imaginary part) on the variables w_j:

$$\begin{aligned} 1 &= \sum_{j=1}^{N} w_j A_j \cos(\varphi_j) \\ 0 &= \sum_{j=1}^{N} w_j A_j \sin(\varphi_j), \end{aligned} \tag{6.7}$$

Equivalently, in a matrix form, we obtain:

$$\mathrm{D}w = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{6.8}$$

where D, named *decodability* matrix, is a $2 \times N$ matrix containing the real and the imaginary parts of the Fourier amplitude of oscillations:

$$\mathrm{D} = \begin{pmatrix} A_1 \cos(\varphi_1) & A_2 \cos(\varphi_2) & \dots & A_N \cos(\varphi_N) \\ A_1 \sin(\varphi_1) & A_2 \sin(\varphi_2) & & A_N \sin(\varphi_N) \end{pmatrix}. \tag{6.9}$$

Because of the Rouché-Capelli theorem for undetermined systems, Eq. 6.8 admits at least one solution if the decodability matrix is full-rank (rank two in the present case of a single frequency target). By expliciting the values of the amplitudes and phases A_i and ϕ_i, one can check that this is the case for every connectivity matrix χ_{ij} which admits at least a pair of non identical eigenvalues. In the only pathological case of a matrix of real and identical eigenvalues, the single neurons response synchronize. Equivalently, the complex amplitudes of the N oscillatory traces are represented in the complex plane by parallel vectors. As a consequence, their real and imaginary part cannot be combined together as in Eq. 6.7 to reconstruct target functions of arbitrary phases.

In the following, we will focus on the well-behaved case of finite-size random reservoirs, which typically admit complex and distinct eigenvalues. In this case, only two free values of $\{w_i\}$ are needed to satisfy the auto-encoder problem for a single frequency target function.

6.1.2 The effective dynamics

We suppose now to select one of the possible solutions. The simplest choice corresponds to fix all the decoding weights to 0 except for the first two. For a particular realization of $g\chi_{ij}$, A_1 and A_2, φ_1 and φ_2 can be computed numerically; once they have been fixed, the solution $\{w_1, w_2\}$ can be derived by solving Eq. 6.6.

We can then close the loop and transform the auto-encoder back into a feedback architecture. The resulting autonomous dynamics is governed by the connectivity matrix $J_{ij} = g\chi_{ij} + u_i w_j$. Such a network admits a self-consistent oscillating solution. We conclude that the eigenspectrum of J_{ij} needs to include a pair of complex conjugate eigenvalues whose real part lies exactly on the critical line, while their imaginary part corresponds to the forcing frequency ω_0. We indicate those eigenvalues with λ_0 and λ_0^*. Note that the value of all the other eigenvalues is not determined. If their real part is smaller than unity, the trained networks is in a marginally stable oscillating state.

We test this procedure in finite networks and we confirm the expectations (Fig. 6.2). In every trial, tuning w^1 and w^2 translates into a pair of eigenvalues with fixed frequency tuned

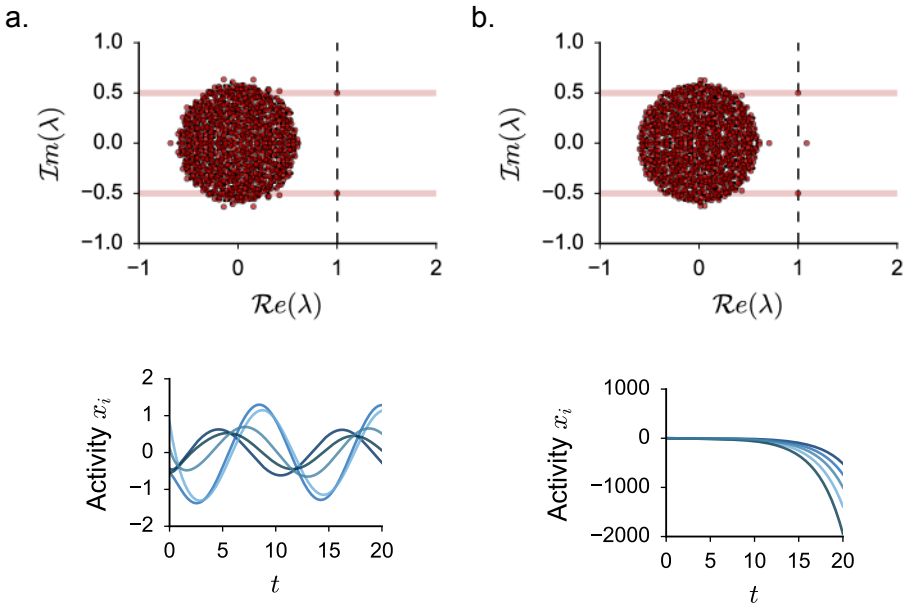


FIGURE 6.2: Transforming auto-encoder solutions into global dynamics. **a.** In this trial, the auto-encoder solution is marginally stable. Top: in red, eigenspectrum of the effective connectivity matrix $J_{ij} = g\chi_{ij} + u_i w_j$. A pair of complex eigenvalues has been tuned to the instability line at 1. Their imaginary part lies exactly on the two horizontal lines where $\mathcal{I}m(\lambda) = \pm\omega_0$. Black dots in the background: original eigenspectrum of $g\chi_{ij}$. Bottom: sample of activity from the closed-loop dynamics (5 randomly chosen units). **b.** The auto-encoder solution is not stable. Figures as in **a**.

on the instability line. Apart from them, most of the remaining eigenvalues still lie close to the circular region, although they never remain in the same position. The position of few of them deviates significantly, and two possible scenarios are verified.

In cases as in Fig. 6.2 **a**, which correspond to the majority of the trials, all the outliers lie below the instability line. As a result, network activity displays marginally stable oscillations. In cases as in Fig. 6.2 **b**, one or more outliers induce an instability to run away activity. As the position of random outliers typically fluctuates around the circle, unstable solutions are in this framework a small minority.

In conclusion we found that, as expected (see [108] and Chapter 8), the stability of the auto-encoder solution is not trivially ensured when the full feedback architecture is considered. Nevertheless, online training algorithms like recurrent least-square minimization might be able to overcome the instabilities: as the degeneracy in the space of the solutions is huge, they could be able to find at least one configuration corresponding to (marginally) stable oscillations. Diverging activity, indeed, would imply finite error values, which force new updates in every value of the decoding weights.

We compare our predictions with the performances of the FORCE online learning scheme. We started from a random feedback architecture as in Fig. 6.1 **a**, and we trained the decoding weights with a recursive least-square (RLS) algorithm [132].

We find that online training can sample and avoid the unstable solutions, resulting in fast convergence around the marginally stable oscillating state (Fig. 6.3 **a**). During training,

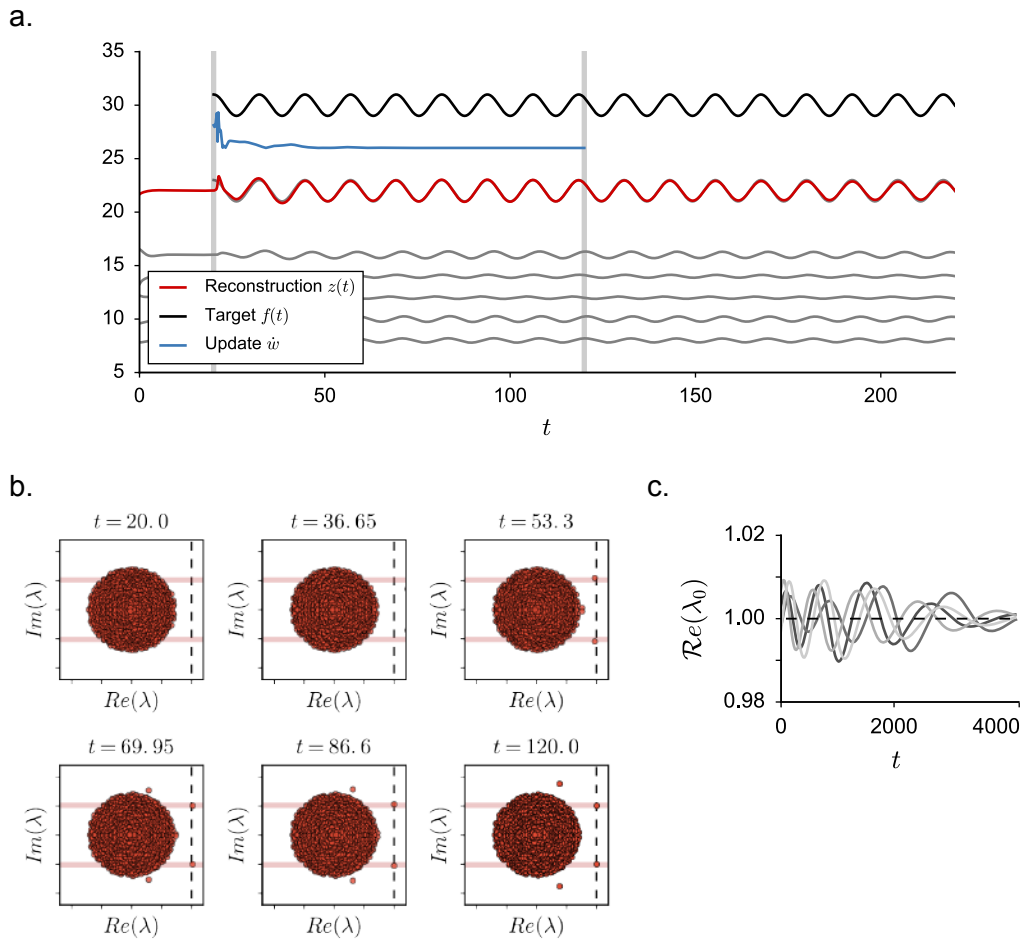


Figure 6.3: Training linear feedback architectures to reproduce a single sinusoid as target function. **a.** Activity before, during and after training through the recursive least-square (RLS) algorithm [132]. Time traces of single units activity are shown in grey. The network reconstruction (in red) well corresponds to the target function (in black) during and after learning. In blue: average size of weight update at each iteration step. As training converges, $\dot{w}$ becomes very small but never equals 0. Choice of the training parameters: learning rate parameter $\alpha = 1$ [132], learning time bin $\delta t = 0.05$ (in units of the implicit network time scale $\tau = 1$). **b.** Eigenspectrum of the effective matrix $J_{ij} = g\chi_{ij} + u_i w_j$ at 6 time points during training. On the pink horizontal lines: $\mathcal{I}m(\lambda) = \pm\omega_0$. **c.** Exact value of the real part of the λ_0 eigenvalue during a long training session. Different traces correspond to different networks realizations. Fluctuations around the critical boundary are very small in size. If the time step of training is small enough, oscillations slowly converge to 1.

the eigenspectrum of the global connectivity matrix is shaped to obtain a pair of marginally stable eigenvalues which correspond to the target frequency (Fig. 6.3 **b**). Because of the finite precision of numerics, however, training convergence is never exact. The least-square algorithm keeps on tuning the synaptic weights, resulting in very small oscillations of the real part of the outliers λ_0 and λ_0^* around the critical value (Fig. 6.3 **b-c**). The fine tuning to the marginally stable state can be improved by adopting smaller time intervals for the weights

update.

If training is precise enough, post-training oscillations will mimic the target function $f(t)$ with high precision for a reasonable time length (as in Fig. 6.3 **a**), before exploding or decaying back to 0.

We notice that, in contrast to the simple solutions that we explicitly designed, the distribution of the decoding weights w_i which are found by FORCE is broad, and it is well fitted by a Gaussian bell (not shown). This comes as a consequence of the step-by-step optimization process which, to begin with, projects the decoding weights in the direction of the random initial condition: $w_i \leftarrow x_i(t_1)$.

To conclude, we remark that training performance in linear networks strongly depends on the values of the reservoir parameters, and it can be compromised by adopting more complex target functions. A more extensive discussion on the topic is presented in the next sections.

6.1.3 Multiple frequencies

As a next stage, we ask whether our feedback architecture is able to learn a target function $f(t)$ which oscillates at more than one frequency. We take for simplicity $f(t) = a_1 \cos(\omega_1 t + \alpha_1) + a_2 \cos(\omega_2 t + \alpha_2)$.

The response of the network consists of two terms which sum linearly:

$$x_i(t) = A_i^1 \cos(\omega_1 t + \varphi_i^1) + A_i^2 \cos(\omega_2 t + \varphi_i^2) \tag{6.10}$$

while the decoder step translates into:

$$a_1 \cos(\omega_1 t + \alpha_1) + a_2 \cos(\omega_2 t + \alpha_2) = \sum_{i=1}^{N} w_i \left[A_i^1 \cos(\omega_1 t + \varphi_i^1) + A_i^2 \cos(\omega_2 t + \varphi_i^2) \right] \tag{6.11}$$

which is now effectively imposing 4 constraints on the N decoding weights w_i.

By induction, one would conclude that the auto-encoder system can be trained on to a maximum number of frequencies which equals $N/2$. In practice, for any target frequency, the network should rely on a pair of oscillators which are not in phase. Loosely speaking, asynchrony is required to balance the two phases in order to align the response with the target phase α_n. As a consequence, in order to learn complex periodic functions, we crucially need the neural response to spread over a wide spectrum of phases.

In the exact solution we derived, the phases and the amplitudes of single neuron activity non trivially depend on the single realization of the system through its eigenvectors and through the weights u_i. As a consequence, as we increase the system size N to accommodate more and more frequencies, it is not straightforward to predict how the phases will distribute.

6.2 A mean field analysis

We thus consider again the easiest case of a single-frequency forcing function $f(t)$ and we ask: what is the limit distribution for the phases of the different units response, in the limit of large networks? Does network activity synchronize in the thermodynamic limit?

In order to answer this question, we develop an alternative description based on a statistical characterization of the amplitudes and the phases in the network response.

For mathematical convenience, we put ourselves in the worst case scenario, where the forcing input is exactly equal for every neuron: $u_i = 1 \; \forall i$. Broadly distributed input weights can nevertheless easily be included in the mean field description.

We aim at providing an effective description of the dynamics:

$$\dot{x_i}(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij} x_j(t) + \cos(\omega_0 t) \tag{6.12}$$

in the limit $N \to \infty$. We inject our guess: $x_i(t) = A_i \cos(\omega_0 t + \phi_i)$ into the equation. We obtain:

$$-A_i \omega_0 \sin(\omega_0 t + \phi_i) = -A_i \cos(\omega_0 t + \phi_i) + \sum_{j=1}^{N} J_{ij} A_j \sin(\omega_0 t + \phi_j) + \cos(\omega_0 t). \tag{6.13}$$

We expand the sinusoidal functions, and we treat separately the term corresponding to $\sin(\omega_0 t)$ ad $\cos(\omega_0 t)$, obtaining:

$$\begin{aligned} -A_i \omega_0 \sin\phi_i &= -A_i \cos\phi_i + \sum_{j=1}^{N} J_{ij} A_j \cos\phi_j + 1 \\ -A_i \omega_0 \cos\phi_i &= A_i \sin\phi_i - \sum_{j=1}^{N} J_{ij} A_j \sin\phi_j. \end{aligned} \tag{6.14}$$

We now define: $X_i = A_i(\cos\phi_i - i \sin\phi_i) = A_i e^{-i\phi_i}$ corresponding to the Fourier amplitude of activity $x_i(t)$ in the frequency ω_0. We define the coupling noise: $Z_i = \eta_i^x + i\eta_i^y$, with $\eta_i^x = \sum_j J_{ij} A_j \cos\phi_j$ and $\eta_i^y = -\sum_j J_{ij} A_j \sin\phi_j$, such that $Z_i = \sum_j J_{ij} X_j$. The term *noise* is adopted from the Dynamical Mean Field (DMF) terminology (see in Chapter 2), and stems from the expectation that coupling terms, which consist of large sums of disordered terms, effectively behave as completely random variables.

By summing the two equations in Eq. 6.14, one finds:

$$X_i = \frac{1 + Z_i}{(1 - i\omega_0)}. \tag{6.15}$$

The solution in the Fourier space is thus given by the sum of two contributions. The first term, $X^0 = 1/(1 - i\omega_0)$, is homogeneous, and coincides with the response that would be measured in non-coupled reservoirs ($g = 0$). The second one, δX_i, is an interaction term, and is proportional to the coupling noise Z_i.

We aim at computing a probability distribution for the real and the complex part of X_i by averaging over different network units, or, equivalently, over different realizations of the random connectivity matrix. Similarly to standard DMF techniques (see in Chapter 2), we start by considering the distribution of the coupling noise Z_i, which can be computed self-consistently. Under the hypothesis that activity decorrelates in very large networks, we obtain:

$$[Z_i] = g \sum_{j=1}^{N} [\chi_{ij}][X_j] = 0 \tag{6.16}$$

as in the random reservoir $[\chi_{ij}] = 0$. This immediately suggests that $[\delta X_i] = 0$.

As all the cross-terms vanish, we are left with three second order statistics to compute:

$$[Z_i^2] = g^2 \sum_{j=1}^{N} \sum_{k=1}^{N} [\chi_{ij}\chi_{ik}][X_j X_k] = g^2 [X_i^2] = g^2 [(X^0 + \delta X_i)^2] = g^2 \left\{ \frac{1 + [Z_i^2]}{(1 - i\omega_0)^2} \right\} \tag{6.17}$$

as $[\chi_{ij}\chi_{ik}] = \delta_{jk}/N$. In the same way we get:

$$\begin{aligned} [Z_i^{*2}] &= g^2 \left\{ \frac{1 + [Z_i^{*2}]}{(1 + i\omega_0)^2} \right\} \\ [|Z_i|^2] &= g^2 \left\{ \frac{1 + [|Z_i|^2]}{1 + \omega_0^2} \right\}. \end{aligned} \tag{6.18}$$

One can easily solve the equations above, finally obtaining the statistics of the interaction noise:

$$\begin{aligned} [Z_i^2] &= \frac{g^2}{(1 - i\omega_0)^2 - g^2} \\ [Z_i^{*2}] &= \frac{g^2}{(1 + i\omega_0)^2 - g^2} \\ [|Z_i|^2] &= \frac{g^2}{1 + i\omega_0^2 - g^2} \end{aligned} \tag{6.19}$$

from which the statistics of the interaction response δX_i can be derived:

$$\begin{aligned} [\delta X^2] &= \frac{g^2}{(1 - i\omega_0)^2} \frac{1}{(1 - i\omega_0)^2 - g^2} = \alpha(g, \omega_0) \\ [|\delta X|^2] &= \frac{g^2}{1 + \omega_0^2} \frac{1}{(1 + \omega_0)^2 - g^2} = \beta(g, \omega_0). \end{aligned} \tag{6.20}$$

Note that we omitted the subscript i, as the population is homogeneous on average.

We finally isolate the real and the imaginary part of the second order statistics:

$$\begin{aligned} [\delta X_x^2] &= \frac{1}{2} (\beta + \mathcal{R}e(\alpha)) \\ [\delta X_y^2] &= \frac{1}{2} (\beta - \mathcal{R}e(\alpha)) \\ [\delta X_x \delta X_y] &= \frac{\mathcal{I}m(\alpha)}{2} \end{aligned} \tag{6.21}$$

where a little algebra gives:

$$\begin{aligned} \mathcal{R}e(\alpha) &= \frac{g^2 \left[1 - 6\omega_0^2 + \omega_0^4 + g^2(\omega_0^2 - 1)\right]}{(1 + \omega_0^2)^2 \left[g^4 + 2g^2(\omega_0^2 - 1) + (1 + \omega_0^2)^2\right]} \\ \mathcal{I}m(\alpha) &= -\frac{2g^2\omega_0(-2 + g^2 + 2\omega_0^2)}{(1 + \omega_0^2)^2 \left[g^4 + 2g^2(\omega_0^2 - 1) + (1 + \omega_0^2)^2\right]}. \end{aligned} \tag{6.22}$$

To sum up, we computed the first and second order statistics of the network activity distribution in terms of its Fourier transform corresponding to the frequency ω_0. We found that, in

the complex plane, the distribution of X_i is centered around the uncoupled response X_0. We examined the statistics of the deviations from X_0 and we found that their real and imaginary parts are correlated. As a consequence, according to a mean field approximation, δX_i must be distributed in the polar plane according to a multi-variate Gaussian law, of mean $(0, 0)$ and covariance matrix defined by Eq. 6.21.

To every point in the complex plane corresponds, in the time domain, one oscillation of frequency ω_0 with fixed phase and amplitude. Instead of trying to compute an explicit joint probability distribution for the amplitudes and the phases, we study the mean field solutions in the polar plane, and we map them numerically in the time domain.

6.2.1 Results

Fig. 6.4 **a** shows the predicted distribution of the Fourier amplitude X_i for fixed values of the parameters g and ω_0. We compare the theoretical expectation with the distribution obtained by numerically simulating activity in a finite-size network ($N = 2000$, right panel) and we find a good agreement.

The fact that the real and the imaginary component of δX_i are correlated implies that, even in disordered networks ($g > 0$), the response in the complex plane can potentially align, thus yielding completely synchronized oscillations.

The exact shape of the predicted distribution depends on both parameters g and ω_0. In general, if we fix ω_0 and we vary g from 0 to 1, we observe a gradual increase in the phase spread of X_i as the interactions become stronger and stronger (Fig. 6.4 **b**). When g is small, the response X_i has small amplitude and is almost aligned on the same phase angle. When g is close to 1, instead, the response is highly delocalized around the common term X_0. We conclude that, for fixed values of ω_0, $g = 1$ represents the optimal parameter choice.

The phase spread shows an additional, non monotonic, dependence on the forcing frequency ω_0. We rigorously quantify the phase spread by measuring the standard deviation of the phases distribution after having remapped them to the interval $[0, \pi]$ (responses with phases γ and $\gamma + \pi$ are considered synchronized as weights w_i can take both positive and negative values). Fig. 6.4 **c** reveals that, when the frequency is too large, the phase spread decreases. For any value of g, furthermore, we find an optimal finite frequency $\tilde{\omega}_0$ for which the phase spread is maximum. The optimal frequency is typically small, it varies non-monotonically with g, and it converges to zero close to the instability in $g = 1$.

We conclude that the maximum phase spread can be achieved with extremely slow target functions and strongly coupled networks ($g \sim 1$). In those conditions, the degeneracy of the solutions to the auto-encoder problem is expected to be large. On the contrary, we found that in large networks complete network synchronization is expected to occur only in completely decoupled reservoirs, which correspond to $g = 0$.

6.3 A comparison with trained networks

Strong desynchronization is expected to be an essential requirement for learning complex periodic functions. In the case of simple sinusoids, however, we predicted that any small but finite phase spread should ensure the existence of a finite number of auto-encoder solutions. These solutions indeed exist and can be computed numerically, as in paragraph 6.1.2, for any value of the parameters ω_0 and g. Numerically computing the eigenspectra of the global con-

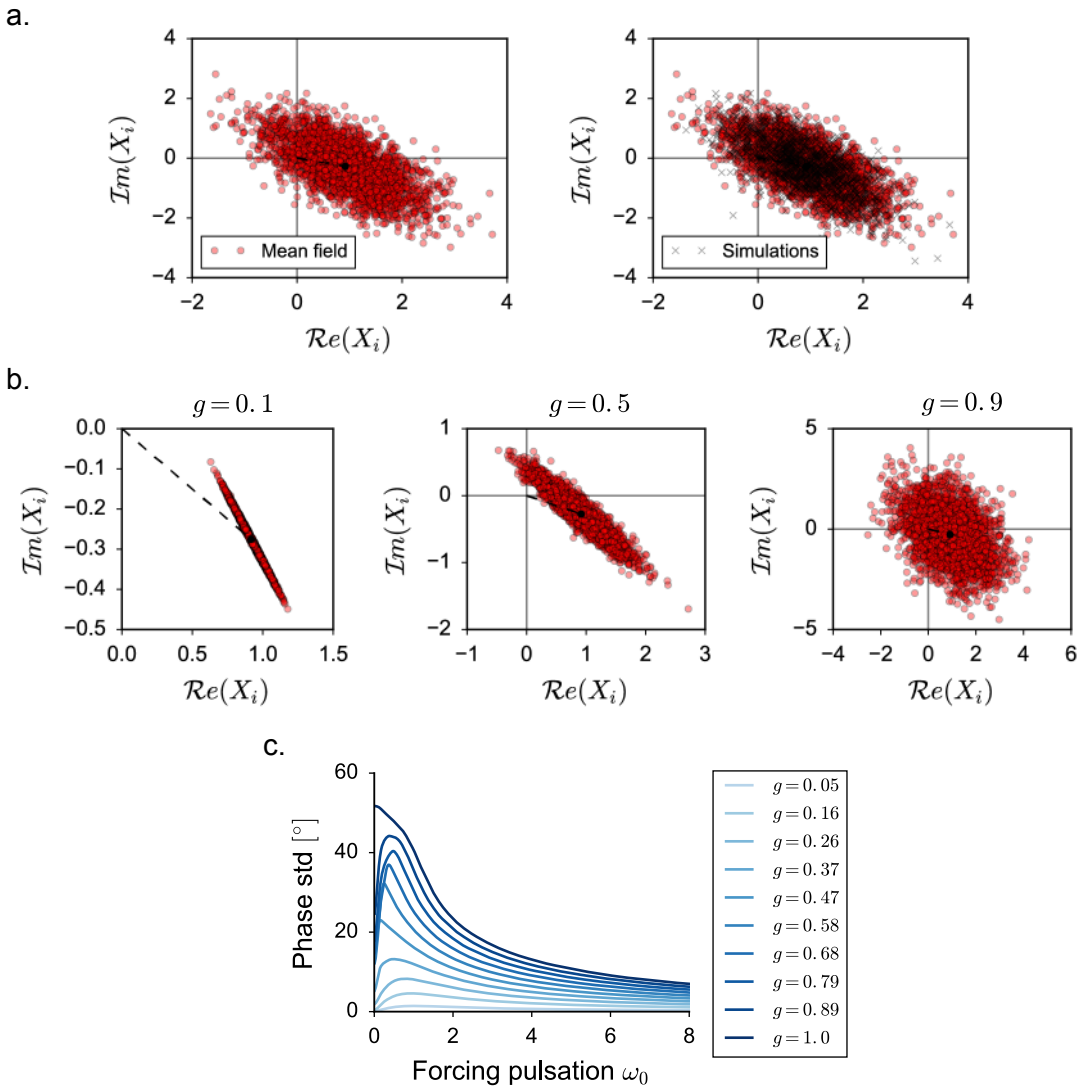


FIGURE 6.4: Network response predicted by the mean field theory: distribution of the Fourier amplitude X_i in the complex plane. **a.** On the left: the theoretically predicted distribution is centered in X_0 (central black dot), and deviations are distributed according to a multivariate Gaussian. Choice of the parameters: $g = 0.8$, $\omega_0 = 0.3$. On the right: comparison with simulated activity in a finite network, $N = 2000$. **b.** Predicted distribution for fixed $\omega_0 = 0.3$ as the coupling g is increased. **c.** Spread in the distribution of the response phases. Networks units are maximally desynchronized when the forcing frequency is small and the coupling strength is high.

nectivity matrix J_{ij} suggests furthermore that a large fraction of them are dynamically stable (see below).

Similarly, we expect the auto-encoder problem to admit at least one solution for any random reservoir which consists of more than two units, as complete synchronization in very small networks is highly implausible to occur.

A different question is whether a batch or an online training algorithm can robustly build a stable solution starting from a small number of almost in-phase oscillators, or whether a wider set of decorrelated basis functions is required.

In order to address this point, we first test online algorithms on the simpler problem of looking for auto-encoder solutions by solving for $\{w_j\}$ the decoding equation $f(t) = \sum_j w_j \phi(x_j)$.

6.3.1 Training auto-encoders

In this setting, the forcing input is clamped to the target function $f(t)$ before, during and after learning. Any change in the decoding weights do not thus propagate to the network dynamics. We use the recursive least square (RLS) algorithm on the weights w_i to enforce numerically and dynamically the decoding step that we solved analytically in the previous section.

As we aim at comparing the results with training performance in full feedback architectures, we measure learning performance by computing the average post-learning reconstruction error within a small time window. As in linear networks the optimal solution is only marginally stable, every successfully trained network suffers indeed of long term stability issues. By measuring post-training error in a relatively short time window we aim at quantifying whether a marginally stable solution was found, while partially ignoring how much fine tuning was achieved.

In Fig. 6.5 **a**, we train the random reservoir on a slow target function ($\omega_0 = 1$), and we look at the average performance as the coupling g is increased. The results show that training performance is seriously impaired at low g values, where oscillating activity is almost synchronized, while it rapidly improves at larger coupling strengths. Furthermore, learning performs better in larger size networks.

In Fig. 6.5 **b-c**, we consider large reservoirs, and we ask how training performance changes when the target pulsation ω_0 is increased. Again, we find that larger target frequencies, and thus larger synchronization, translate into more serious training impairments.

6.3.2 Training feedback architectures

In a second step, we apply the RLS algorithm to the original feedback architecture as in [132]. In this configuration, the forcing input $z(t)$ is kept dynamically close to the target function $f(t)$ through the online weights update.

We find that training the whole feedback architecture typically results in worse performances. This suggests that, even if a solution exists and can be found algorithmically starting from the asymptotic network response, an addition effort is required to efficiently stabilize the recurrent feedback dynamics.

However, similarly to what observed in Fig. 6.5 in the case of clamped inputs, the error decreases in the parameter regions which imply larger desynchronization in the reservoir response.

More in detail, in the left panel of Fig. 6.6 **a**, we show that the post-training reconstruction error sensibly decreases as g increases runs from 0 to 1.

As in this framework the training algorithm is searching for full feedback solutions, one could hypothesize that failures at small g derive from an increase in the fraction of unstable solutions. To rule out this possibility, we estimate the relative number of unstable solutions by solving numerically the decoding equations as in paragraph 6.1.2. We iterate this proce-

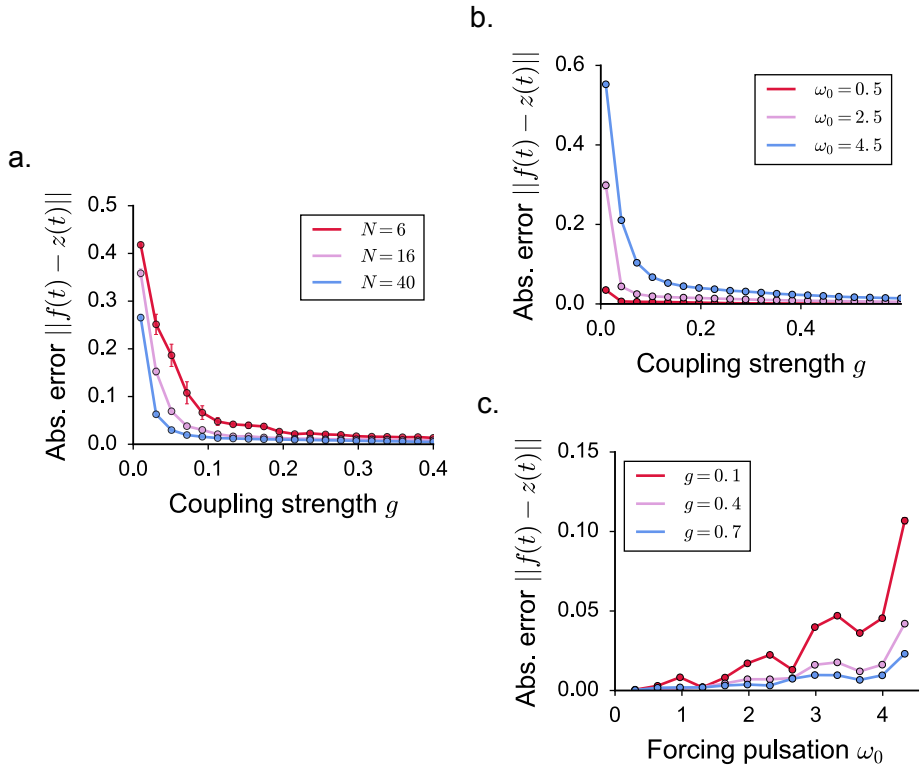


FIGURE 6.5: Recursive least square (RLS) performance in training the decoding weights w_i when the reservoir input is clamped to the target function $f(t) = cos(\omega_0 t)$. Training parameters as in Fig. 6.3. Performance is measured as the average absolute error within one time step of integration of post-learning activity, during a short time interval T after training ($T = 30$). Average over 15 different trained networks. As the target frequency is increased, smaller training time steps have been adopted. **a.** Average reconstruction error in small networks for increasing values of the coupling g. Even if one solution exists for any network size $N > 2$, training performance depends on the network size. Performance, furthermore, significantly improves in strongly coupled networks, as the oscillating response desynchronizes. Note that extremely small values of the learning rate parameter α seem to alleviate training impairments at small g values, but they don't prevent large errors in the post-learning reconstruction signal. Forcing pulsation: $\omega_0 = 1$. **b-c.** Performance in larger networks ($N = 300$) as a function of g for faster and slower target functions. Details as in **a**. At smaller frequencies, activity is less synchronized and learning is more precise.

dure over many trials and we assess stability by looking at the eigenspectrum of the global connectivity matrix $J_{ij} = g\chi_{ij} + u_i w_j$.

We find that the fraction of unstable solutions is extremely small at small g values, while it approaches one half when the eigenvalues circle is close to the instability line (Fig. 6.6 **a**, right panel). The fraction of unstable solutions grows almost linearly with g and does not significantly depend on the networks size N and on the target frequency ω_0. At high g values, the portion of unstable solutions is high but the reconstruction error is small in every training trial. We conclude that, at least in the case of simple target functions and strong reservoir desyn-

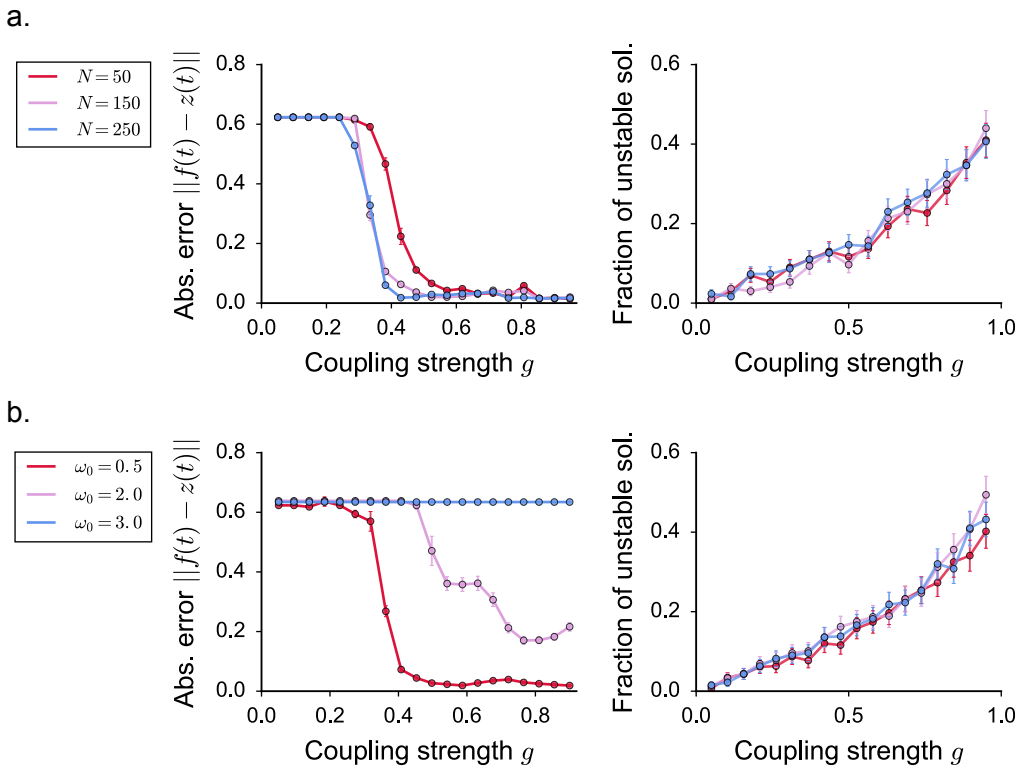


Figure 6.6: RLS algorithm performance in training the decoding weights when the feedback input to the reservoir $z(t)$ is constructed dynamically from network activity. Training parameters and details as in Fig. 6.5. **a.** Training feedback architectures for different values of the coupling strength g and different reservoir sizes N. Left panel: training performance measured as the average post-training reconstruction error. Right: brute force estimation of the fraction of unstable open-loop solutions. Training failures at small g values cannot be due to an increase in the fraction of unstable solutions, as the trend goes in the opposite direction. Forcing pulsation: $\omega_0 = 0.5$. **b.** Training feedback architectures for different values of the coupling strength g and different target pulsations ω_0. Details as in **a**, $N = 150$.

chronization, online training algorithms can robustly isolate stable from unstable solutions.

In Fig. 6.6 **b** we show that, similarly to the auto-encoder configuration, training improves dramatically when adopting slow target functions. In the right panel we show that this effect cannot be explained by instability arguments either.

6.3.3 Discussion

To sum up, we found that the common RLS algorithm is not always able to stably converge to a solution, even in cases where more than one solution is analytically guaranteed to exist. Training performance improves significantly in the parameter regions where the mean field theory predicts the phase spread to be maximal. In these phase space regions, training algorithms can rely on a widely distributed basis set of oscillations which can be efficiently balanced to align the network reconstruction to the target response.

Training failures cannot be due to dynamical instabilities, as they also occur in the open-loop setting. They more probably derive from severe precision limitations in the numerical computations that are prescribed by the recursive least-square (RLS) algorithm. In this hypothesis, synchronized reservoir activity and ill-conditioned computations would emerge, respectively, as the dynamical and statistical expressions of the same fundamental problem.

In order to test this hypothesis in a theoretically simplified framework, one could try to solve numerically the whole decodability set of equations (Eq. 6.8). Similarly to the RLS algorithm, one can consider its least square solution, which is given by:

$$w = \mathrm{D}^T(\mathrm{D}^T\mathrm{D})^{-1}F. \tag{6.23}$$

Our analytical approach ensures the square matrix $\mathrm{D}^T\mathrm{D}$ to be full rank and thus invertible for large and random reservoirs. It does not exclude, however, that the matrix might be very ill-conditioned for certain values of the network parameters. Note that the entries of the $\mathrm{D}^T\mathrm{D}$ matrix are naturally returned by our mean field approach, so that further analytical steps might be easily taken in this direction. We leave those investigations to future analysis.

6.4 Towards non-linear networks

The numerical analysis we performed suggests that linear networks optimally learn arbitrarily slow signals in a parameter region close to the instability due to large couplings strengths.

In non-linear networks, the overall coupling g can be pushed to large values without inducing destabilizations to run away activity. A different kind of instability is expected to occur at $g = 1$ (see also Chapters 1 and 2), which leads to the onset of irregular chaotic fluctuations. In presence of external forcing inputs, however, the instability coupling is expected to be larger than unity. The exact value of g_C non trivially depends on the amplitude and the frequency of the periodic forcing [99].

As it was already mentioned, a full characterization of non-linear networks response is a challenge that presents several technical difficulties. Here we take an exploratory approach and we compute numerically the phase distribution that we predicted analytically in the case of linear networks (Fig. 6.4 **c**).

6.4.1 Response in non-linear random reservoirs

We consider values of the coupling strengths at which network activity is still deterministic (in the case of our target function: $0 < g \lesssim 1.8$). When a sinusoidal input is injected into a non-linear network, together with a response in the forcing frequency, a response in the harmonic frequencies can be elicited as well. Multi-frequencies responses dominate at high coupling values. We isolate the phase of oscillations at the main pulsation ω_0 through a numerical Fourier decomposition.

The result is shown in Fig. 6.7. To begin with, we observe that the order of magnitude of phase spread for linear and non-linear networks is essentially comparable. As in linear networks, furthermore, the phase dispersion increases with g. For any coupling value, our results reveal a finite optimal frequency at which desynchronization is maximum. For high values of g, sinusoidal functions with pulsation $\omega_0 \sim 1$ (in the implicit network time scale, given by $\tau = 1$) result into an optimal response.

We finally observe that, similarly to what we found in Section 6.2, the phase spread falls down monotonically at large ω_0 values when the coupling is smaller than 1. At high g values,

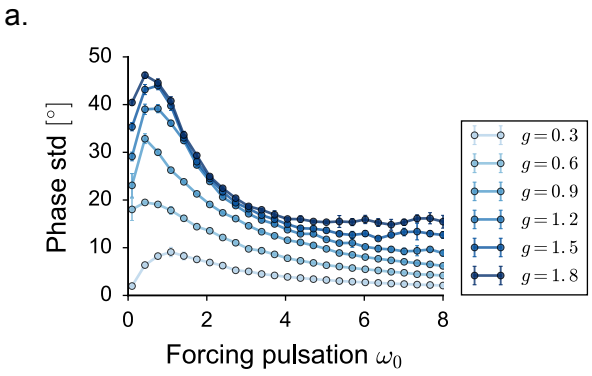


Figure 6.7: Non-linear network response: numerical analysis. **a.** Phase dispersion, measured as the standard deviation of the phase distribution in the support $[0, \pi]$ across different units of the same network. Average over 5 sample networks, $N = 3000$. We restricted the analysis to the response relative to the peak pulsation ω_0.

instead, the degree of synchrony seem to saturate to constant values in the high frequency domain (for $\omega_0 > 4$, it converges around $18°$ when $g = 1.8$), suggesting that non-linear networks can more robustly respond to high frequency forcing in the parameter region where chaotic fluctuations are quenched by the external input.

6.4.2 Training non-linear networks

If synchronization properties are similar in linear and non-linear reservoirs, one can ask whether online training algorithms require strongly desynchronized activity also when training is performed within non-linear architectures. To conclude this chapter, we thus perform a systematic analysis of training performances within non-linear architectures. As it will be shown, the training impairment which is observed at weak coupling strengths strongly resembles the problems that have been observed in linear models.

It has been already observed that the FORCE training performance is optimal for parameter values near to the *edge of chaos*, i.e. at coupling strengths g which are close to g_C [132]. We replicate this result in our framework by training non-linear feedback architectures on a single frequency target function, and we measure performance by computing the after training reconstruction error over a long time window T.

As expected, performance is optimal in the $g > 1$ region where activation traces are strongly non-linear and network activity is stabilized thanks to the readout feedback input (Fig. 6.8 **a** and **c**, third row). Above g_C, activity becomes unpredictable, and chaotic fluctuations bring the network reconstruction far from the target function, resulting in a large average error (Fig. 6.8 **c**, fourth row).

A similarly large error is measured at small g values. In contrast to the high coupling region, however, when $g \lesssim 1$, training correctly converges. For finite training steps, nevertheless, the reconstruction $z(t)$ is never exactly tuned to the target frequency. As a result, when post-synaptic activity is integrated over a long time window, the target and the reconstruction functions completely desynchronize, and the training error is on average large (Fig. 6.8 **c**, first and second row). This kind of error can be controlled and decreased by adopting smaller and

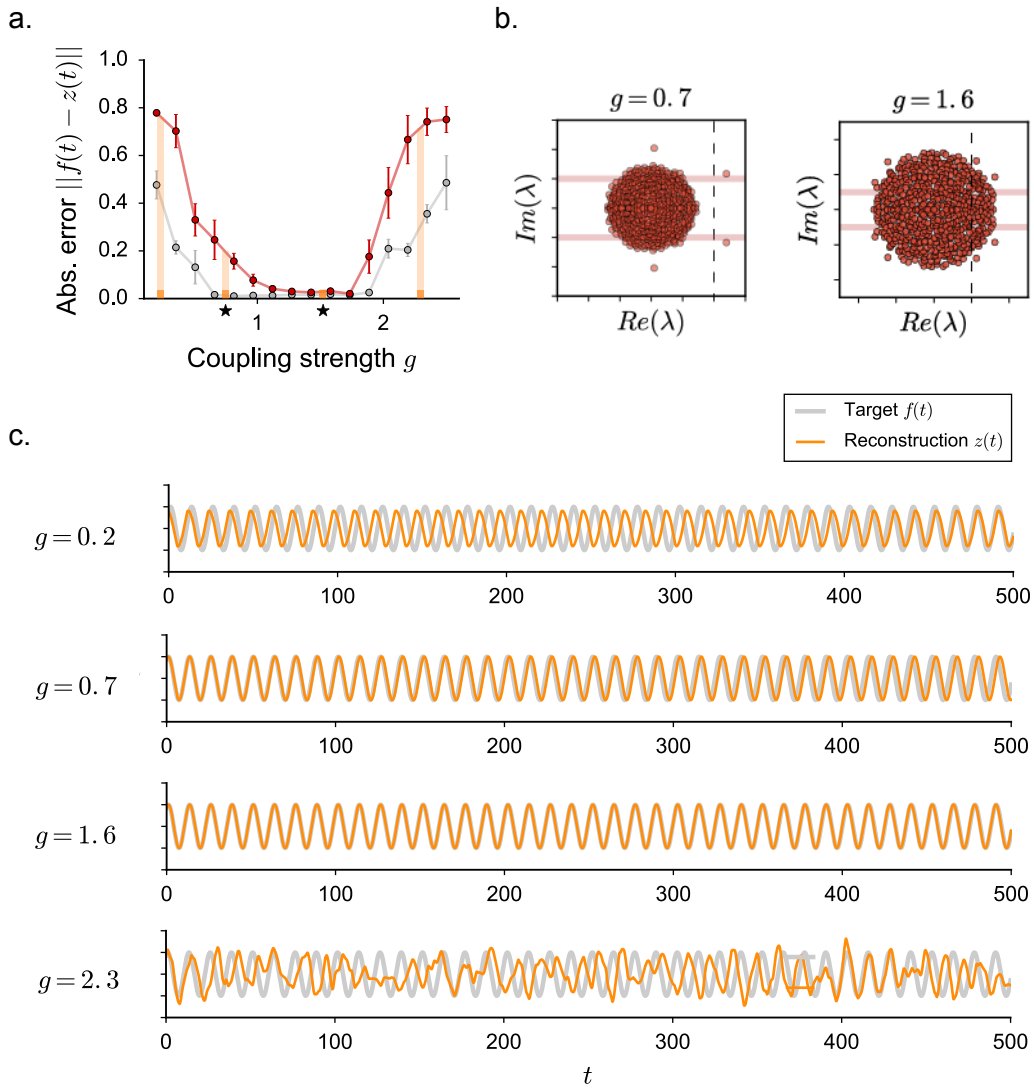


FIGURE 6.8: FORCE training within the original non-linear architecture. Target function: $f(t) = \cos(\omega_0 t)$, $\omega_0 = 0.5$. Training parameters as in Fig. 6.3. **a.** Performance, measured as the average absolute error during post-training reconstruction. Red line: the error is averaged over a long time window: $T = 500$ (in the implicit network time scale where $\tau = 1$). A large reconstruction error is measured similarly in the almost linear ($g < 1$) and chaotic ($g > 1.9$) regimes. Grey line: the error is averaged over a short time window, $T = 30$, comparable with the choice we adopted in the previous sections with linear networks. Note that with this choice, the training performance in the weak coupling regime is quantitatively comparable with the results we obtained by training purely linear feedback architectures. **b.** Eigenspectrum of the global network connectivity $J_{ij} = g\chi_{ij} + w_j$ at the end of training. We considered two values of the synaptic strength g, indicated in **a** by the black stars. On the pink horizontal lines: $\mathcal{I}m(\lambda) = \pm\omega_0$. **c.** Target function (grey) and network reconstruction (orange) after the end of the learning period. We adopted four values of g, which are indicated in **b** by the orange bars. While the long-term reconstruction error is large both at small and large g values, in the weak coupling regime the error comes from a small frequency mismatch.

smaller update time steps. Note that in non-linear networks, oscillating activity appears as a result of a saturation-stabilized Hopf bifurcation. As a result, tuning both the real and the imaginary part of the eigenvalues lying above the critical line directly affects the final frequency of oscillations.

The same behaviour has been observed in linear networks, where the online tuning of the forcing eigenvalues λ_0 and λ_0^* is never exact, and results in a mismatch (together with decaying or exploding activity) on the long run.

Learning in the weak coupling regime presents also other similarities with training in the linear case. First, the training performance smoothly improves with increasing coupling values (Fig. 6.8 **a**). After training, furthermore, the fixed point in 0 is unstable only in one planar direction, which is defined by two complex conjugate eigenvectors whose eigenvalues lie above the critical line (Fig. 6.8 **b**, left). Their imaginary part reasonably well explains the learnt target frequency. Finally, in the weak coupling regime, training acts through very strong weights modifications during a first phase, and it keeps on tuning the exact value of the unstable eigenvalues during the second phase (not shown).

As g becomes close to 1, the global dynamics properties start to diverge smoothly from almost-linear features. The fixed point in zero becomes unstable with respect to a large number of directions, and a direct check at its stability matrix (coinciding with J_{ij}, as $\phi'(0) = 1$), is no longer very informative (Fig. 6.8 **b**, right). The eigenspectrum that we obtain, furthermore, changes significantly and quantitatively in different training trials.

In those conditions, learning is extremely stable over time and the error is minimized. One possibility would be that training in this regime is taking advantage of higher and higher desynchronization properties which increase monotonically with g (see in Fig. 6.7). However, the complete picture is likely to be more complex. Our analysis, indeed, did not take into account several different elements which are likely to play an important role in strongly connected non-linear networks. Above all, we did not consider the effect of having multiple responses at higher harmonics, and we did not address the consequences of a significative change in the phase space topology, suggested by the large number of unstable directions for the stationary solution in 0.

How those different elements contribute to extremely high precision training is still an open question.

Part III

LINKING CONNECTIVITY, DYNAMICS AND COMPUTATIONS

Summary of Chapters 7 - 8 - 9

Synaptic connectivity determines the dynamics and computations performed by neural circuits. Due to the highly recurrent nature of circuitry in cortical networks, the relationship between connectivity, dynamics and computations is complex, and understanding it requires theoretical models. Classical models of recurrent networks are based on connectivity that is either fully random or highly structured, e.g. clustered. Experimental measurements in contrast show that cortical connectivity lies somewhere between these two extremes. Moreover, a number of functional approaches suggest that a minimal amount of structure in the connectivity is sufficient to implement a large range of computations.

Based on these observations, here we develop a theory of recurrent networks with a connectivity consisting of a combination of a random part and a minimal, low dimensional structure. We show that in such networks, the dynamics are low-dimensional and can be directly inferred from connectivity using a geometrical approach. We exploit this understanding to determine minimal connectivity structures required to implement specific computations. We find that the dynamical range and computational capacity of a network quickly increases with the dimensionality of the structure in the connectivity. Our simplified theoretical framework captures and connects a number of outstanding experimental observations, in particular the fact that neural representations are high-dimensional and distributed, while dynamics are low-dimensional, with a dimensionality that increases with task complexity. Altogether our results suggest a simple general principle for relating connectivity, dynamics and computations: the low-dimensional structure of the connectivity matrix determines low-dimensional dynamics and computations in recurrent networks.

A substantial fraction of this part of the dissertation is adapted from the manuscript: *Linking dynamics, connectivity and computations* by F. Mastrogiuseppe and S. Ostojic, submitted.

Dynamics of networks with unit rank structure 7

Understanding the relationship between synaptic connectivity, neural activity and behavior is the central endeavor of neuroscience. Networks of neurons encode incoming stimuli in terms of electrical activity, and transform this information into decisions and motor actions through synaptic interactions, thus implementing computations that underly behavior. Reaching a simple, mechanistic grasp on the relation between connectivity, activity and behavior is however highly challenging. Cortical networks, which are believed to constitute fundamental computational units in the mammalian brain, consist of thousands of neurons that are highly inter-connected through recurrent synapses. Even if one was able to experimentally record the activity of every neuron and the strength of each synapse in a behaving animal – the ultimate goal of current technological developments –, understanding the causal relationships between these quantities would remain a daunting task because an appropriate conceptual framework is currently lacking [117, 51]. Simplified, computational models of neural networks provide a testbed for developing such a framework. In computational models, the strengths of all synapses and the activity of all neurons are known, yet an understanding of the relation between connectivity, dynamics and computations has been achieved only in very specific cases [17, 29, 124, 147, 81].

One of the most popular and best-studied classes of network models is based on fully random recurrent cortical connectivity [127, 24, 143]. Such networks display self-sustained irregular activity that closely resembles spontaneous cortical patterns recorded *in-vivo* [126, 119, 120]. The relationship between connectivity and dynamics can be understood in great detail in this case, and randomly-connected networks have become a central theoretical paradigm that has led to the development of fundamental concepts such as excitation-inhibition balance and decorrelation [105]. However, randomly connected recurrent networks display only very stereotyped responses to external inputs and can implement only a limited range of input-output computations [99, 74]. To implement more elaborate computations, classical network models rely instead on highly structured connectivity, in which every neuron belongs to a distinct cluster, and is selective to only one feature of the task [147, 9, 81]. Actual cortical connectivity appears to be neither fully random nor fully structured [59, 129], and the activity of individual neurons displays a similar mixture of stereotopy and disorder [107, 83, 32]. To take these observations into account, and implement general-purpose computations in recurrent

networks, a number of functional approaches have been developed for designing appropriate connectivity matrices [65, 67, 80, 132, 48, 13]. A general conceptual picture of how connectivity determines dynamics and computations is however currently missing.

Remarkably, albeit developed independently and motivated by different goals, several of the functional approaches for designing connectivity appear to have reached similar solutions, in which the computations performed by the network rely on a minimal, low rank structure in the synaptic matrix. In classical Hopfield networks [65, 110, 104], a rank one term is added to the connectivity matrix for every item to be memorized. In echo-state [67, 80] and FORCE learning [132], and similarly within the Neural Engineering Framework [48], computations are implemented through feedback loops that are mathematically equivalent to adding rank one components to the otherwise random connectivity matrix. In the predictive spiking theory [20] the requirement that information is represented efficiently leads again to a connectivity matrix with low rank structure. Taken together, the results of these studies suggest that low rank structure added on top of random recurrent connectivity may provide a general and unifying framework for implementing computations in recurrent networks.

Based on this observation, in the last part of this dissertation we develop a theory of large, random networks perturbed by weak, low dimensional connectivity, and examine the computational capacity of such a setup.

We start by showing that in such networks, both spontaneous and stimulus-evoked activity are low dimensional and can be predicted from the geometrical relationship between a small number of high-dimensional vectors that represent the connectivity structure and the incoming stimuli. This understanding of the relationship between connectivity and network dynamics will allow us in the next chapters to directly design minimal, low rank connectivity structures that implement specific computations.

Here, we start by considering the simplest possible type of low dimensional connectivity, a matrix P_{ij} with unit rank. Examples of higher rank structures will be discussed in Chapter 8.

7.1 One dimensional spontaneous activity in networks with unit rank structure

We studied a classical network of N firing rate units with a sigmoid input-output transfer function [127, 132, 73]. The connectivity matrix consisted of a sum of a structured, low rank matrix P_{ij} and a random, full-rank matrix of strength controlled by a parameter g, which we also denote as disorder strength. We considered the low rank component of the connectivity to be fixed and uncorrelated with the random part, which was considered unknown except for its statistics (mean 0, variance $\frac{g^2}{N}$). In absence of structured connectivity, the dynamics are determined by the strength g of the random connectivity: for $g < 1$, the activity in absence of inputs decays to zero, while for $g > 1$ it displays strong, chaotic fluctuations [127]. Our aim was to understand how the interplay between the fixed, low rank part and the random part of the connectivity shapes the dynamics of activity in the network. We first describe spontaneous dynamics in the network, and later turn to effects of inputs.

We found that an effective, statistical description of the dynamics can be mathematically derived if the network is large and the low dimensional part of the connectivity is weak (i.e. if P_{ij} scales inversely with the number of units N in the network). In this situation, the activity of the unit i can be described in terms of the mean and variance of the total input it receives,

determined by averaging over different realizations of the random part of the connectivity matrix. Dynamical equations for these quantities can be derived by extending the classical dynamical mean field theory [127]. Full details of the analysis are provided in Section 7.3, here we summarize the main results.

A matrix P_{ij} with unit rank can be written as:

$$P_{ij} = \frac{m_i n_j}{N} \tag{7.1}$$

where $m = \{m_i\}$ and $n = \{n_j\}$ are two N-dimensional vectors which we call respectively the right- and left-structure vectors. These vectors fully specify the structured part of the connectivity, and we consider them arbitrary, but fixed and uncorrelated with the random part of the connectivity. In the following, we will show that the network dynamics can be directly understood from the geometrical arrangement of the vectors m and n.

Our analysis reveals that at equilibrium, the average input μ_i to unit i is given by

$$\mu_i = \kappa m_i, \tag{7.2}$$

where

$$\kappa = \frac{1}{N} \sum_{j=1}^{N} n_j [\phi_j]. \tag{7.3}$$

The scalar quantity κ represents the overlap between the left structure vector n and the N-dimensional vector $[\phi] = \{[\phi_j]\}$ that describes the average firing activity of the units in the network ($[\phi_j]$ is the firing rate of unit j averaged over different realizations of the random component of the connectivity). The overlap κ therefore quantifies the degree of structure along the vector n in the activity of the network. If $\kappa > 0$, the equilibrium activity of each neuron is correlated with the corresponding component of the vector n, while $\kappa = 0$ implies no such structure is present.

For a given realization of the random component of the connectivity, the equilibrium input to unit i will deviate from the expected mean μ_i, and these static fluctuations can be quantified by the corresponding, static variance. Strong random connectivity may moreover induce chaotic fluctuations, which lead to an additional temporal variance. The overlap κ and the static and temporal variances are macroscopic network quantities that obey a set of coupled equations (see Eqs. 7.17, 7.29 and 7.66). Those equations can be solved to determine the possible regimes of network dynamics.

Similarly to fully random networks, two general types of activity can emerge: static, fixed point dynamics, and fluctuating, chaotic activity. We start by describing static dynamics, expected to occur when the random part of the connectivity is not too strong.

If one represents the network activity as a point in the $N-$dimensional space where every dimension corresponds to the activity of a single unit, Eq. 7.2 shows that the structured part of the connectivity induces a one-dimensional organization of the spontaneous activity along the vector m. This one-dimensional organization however emerges only if the overlap κ does not vanish. As the activity of the network is organized along the vector m, and κ quantifies the projection of the activity onto the vector n, non-vanishing values of κ require a non-vanishing overlap between vectors m and n. This overlap, given by $m^T n/N = \sum_j m_j n_j/N$ corresponds in fact to the eigenvalue of the rank one matrix P_{ij} and directly quantifies the strength of the structure in the connectivity. The connectivity structure strength $m^T n/N$ and the activity structure strength κ are therefore directly related, but in a highly non-linear manner. If the

connectivity structure is weak, the network only exhibits homogeneous, unstructured activity corresponding to $\kappa = 0$, so that the average input is zero for all units (Fig. 7.1 **b** blue). If the connectivity structure is strong, structured, heterogeneous activity emerges ($\kappa > 0$), and the activity of the network at equilibrium is organized in one dimension along the vector m (Fig. 7.1 **b** green and Fig. 7.2 **b**), while the random connectivity induces additional fluctuations along the remaining $N-1$ directions. Note that because of the symmetry in the specific input-output function we use, when a heterogeneous equilibrium state exists, the configuration with the opposite sign is an equilibrium state too, so that the network activity is bistable (for more general asymmetric transfer functions, this bistability is still present, although the symmetry is lost, see Appendix C).

The random part of the connectivity disrupts the organization of the activity induced by the connectivity structure through two different effects. The first effect is that increasing the disorder strength g leads to stronger fluctuations of equilibrium inputs around the average values μ_i (Fig. 7.2 **c**). As the fluctuations are identical for all units (i.e. unstructured), their increase results in a progressive reduction of the structure in the activity quantified by κ (Fig. 7.2 **b**). A second, distinct effect is that increasing the disorder strength tends to destabilize equilibrium activity.

The stability of the dynamics can be assessed by examining the temporal evolution close to equilibrium, which is in general determined by the spectrum of eigenvalues at the corresponding fixed point. In our case, this spectrum consists of two components: a continuous, random component distributed within a circle in the complex plane, and a single outlier induced by the structured part of the connectivity (Fig. 7.1 **b**). The radius of the continuous component and the value of the outlier depend on the connectivity parameters. Although the two quantities in general are non-trivially coupled, the value of the radius is mostly controlled by the strength of the disorder, while the value of the outlier increases with the strength of the rank one structure (Fig. 7.2 **a**). The equilibrium is stable as long as the real part of all eigenvalues is less than unity.

The appearance of one dimensional structured activity with increasing connectivity structure strength corresponds to the instability induced by the outlier crossing unity (Fig. 7.1 **b** green). Increasing the disorder strength on the other hand leads to another instability, corresponding to the radius of the continuous component crossing unity (Fig. 7.1 **b** orange and red). This instability gives rise to chaotic, fluctuating activity. To describe this type of dynamics, the macroscopic equations for network activity need to be supplemented with a variable quantifying temporal fluctuations.

Similarly to static activity, depending on the strength of the structured connectivity two different types of chaotic dynamics can emerge. If the disorder in the connectivity is much stronger than structure, the overlap κ is zero (Fig. 7.2 **b**). As a result, the mean activity of all units vanishes and the dynamics consist of unstructured, $N-$dimensional temporal fluctuations (Fig. 7.2 **c**), as in the classical chaotic state of fully random networks (Fig. 7.1 **b** red). In contrast, if the strengths of the random and structured connectivity are comparable, the overlap κ is non-zero, and a new type of chaotic activity emerges, in which $\kappa > 0$ so that the mean activity of different units is structured in one dimension along the direction m as shown by Eq. 7.2, but the activity of different units now fluctuates in time (Fig. 7.1 **b** orange). Similarly to structured static activity, in this situation the system is bistable as states with opposite signs are always admissible.

The phase diagram in Fig. 7.1 **a** summarizes the different types of dynamics that can emerge as function of the strength of structured and random components of the connectivity

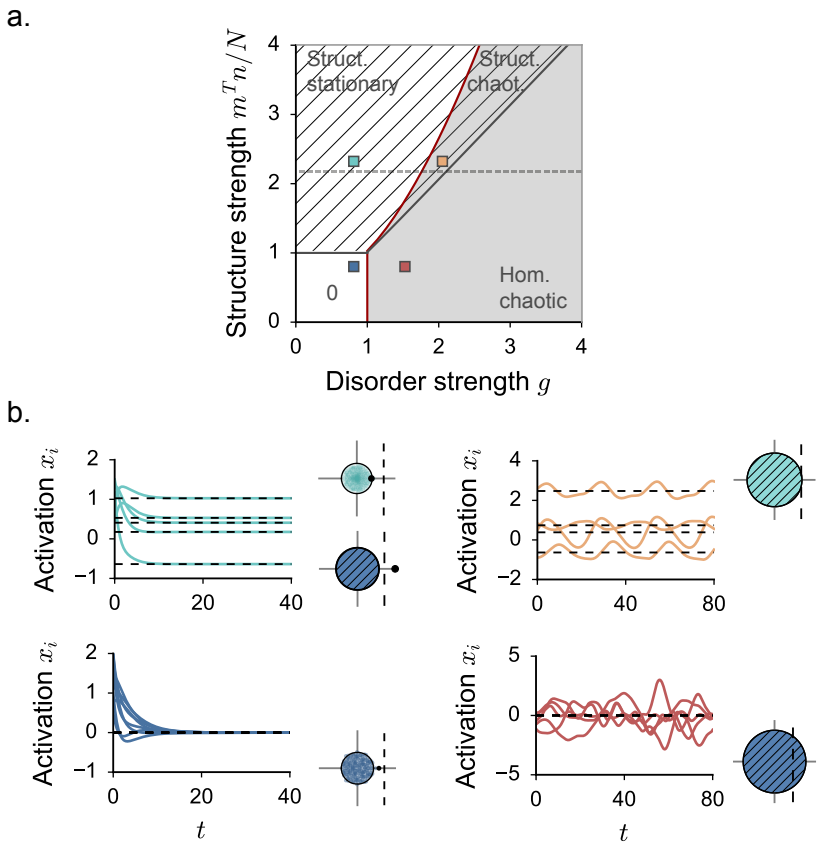


FIGURE 7.1: Spontaneous activity in random networks with unit rank connectivity structure. Results from the Dynamical Mean Field (DMF) theory: phase diagram. **a.** Dynamical regimes of the network activity as function of the structure connectivity strength $m^T n/N$ and the disorder strength g. Hatched areas indicate the parameter regions where two stable DMF solutions exist and network activity is bistable. Shaded areas indicate the phase space regions where network dynamics are chaotic. **b.** For parameter values indicated by the colored dots in the phase diagram, samples of activity from finite networks simulations are shown. Time is renormalized by the time constant of individual units. Next to each panel, the eigenspectrum of the stability matrix of the homogeneous fixed point ($S_{ij} = J_{ij}$) is displayed in blue. When a structured, bistable stationary state exists, the corresponding eigenspectrum ($S_{ij} = \phi'(x_j^0)J_{ij}$) is shown in green. Dots: the eigenvalues are computed numerically if the corresponding state is stable; black lines: theoretical prediction. In this figure, $\Sigma_m = 1.0$ and $\Sigma_n = 0.2$. Note that the precise position of the instability to chaos depends on the value of Σ_m.

matrix. Altogether, the structured component of connectivity favors a one-dimensional organization of network activity, while the random component favors high-dimensional, chaotic fluctuations. Particularly interesting activity emerges when the structure and disorder are comparable, in which case the dynamics show one-dimensional structure combined with high-dimensional temporal fluctuations. This structured chaotic activity can give rise to dynamics with very slow timescales [66] (see Appendix D).

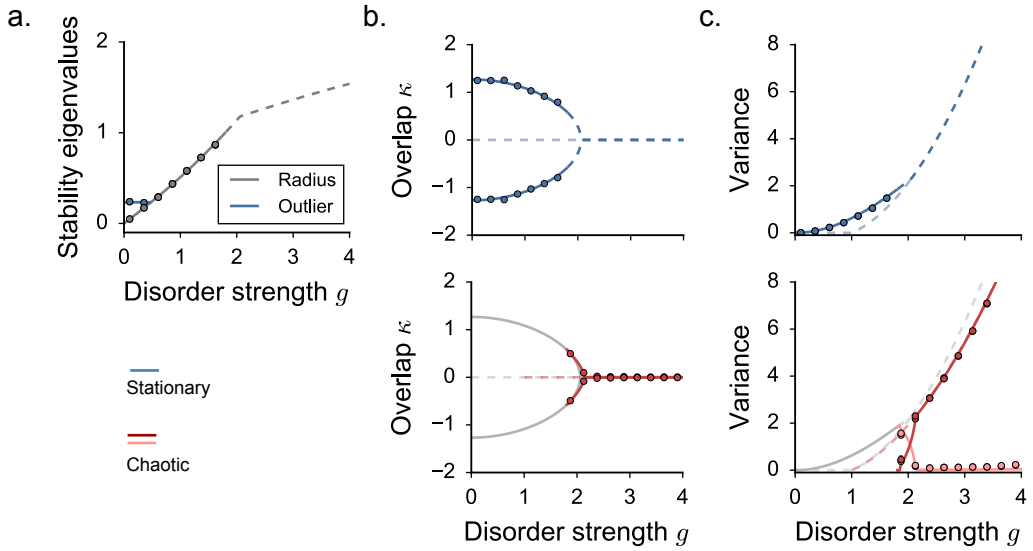


Figure 7.2: Spontaneous activity in random networks with unit rank connectivity structure. Results from the Dynamical Mean Field (DMF) theory: bifurcation diagrams. Network activity stability properties and statistics as the disorder strength g is increased and the structure strength is fixed to 2.2 (dashed horizontal line in the phase diagram in Fig. 7.1 **a**). **a.** Stability of the structured stationary state. Theoretical prediction for the radius of the compact part of the eigenspectrum and the outlier position. **b.** Amount of structure in the activity along the vector m, as quantified by $\kappa = \langle n_i[\phi_i] \rangle$. **c.** Variance of the input to a given network unit induced by random connectivity. Blue and pink: static variance, red: temporal variance. In **b-c**, top panels display statistics for stationary dynamics and bottom panels display statistics for chaotic activity. The solutions of DMF theory are displayed as continuous (resp. dashed) lines if they correspond to a stable (resp. unstable) dynamics. Dots: network activity statistics measured in simulations of finite-size networks, starting from initial conditions centered around m and -m. Activity is integrated up to $T = 800$. In simulations, $N = 5000$, and statistics are averaged over 15 different network realizations. The error bars, when visible, correspond to the standard deviation of the mean (as in every other figure, if not differently specified). The structure vectors m and n were generated from Gaussian distributions, and overlap only along the unitary direction ($M_m > 0$, $M_n > 0$, $\rho = 0$, see Section 7.3). As shown in Section 7.3, qualitatively similar regimes are obtained when the overlap is defined on an arbitrary direction. In this figure $\Sigma_m = 1.0$ and $\Sigma_n = 0.2$. Note that the precise position of the instability to chaos depends on the value of Σ_m.

7.2 Two dimensional activity in response to an input

Having described spontaneous activity in networks with rank one connectivity structure, we now turn to the response to an external input (Fig. 7.3 **a**). Our effective statistical description can be directly extended to that situation, and predicts that if each unit i receives a constant external input I_i, at equilibrium its total input is on average:

$$\mu_i = \kappa m_i + I_i. \tag{7.4}$$

At the level of the N-dimensional space representing the activity of the whole population,

Eq. 7.4 shows that the network activity in presence of an input lies on the two-dimensional plane spanned by the right-structure vector m and the vector $I = \{I_i\}$ that corresponds to the pattern of external inputs to the N units. The contribution of the vector m to this two-dimensional activity is quantified by the overlap κ between the network activity $[\phi]$ and the left-structure vector n introduced in Eq. 7.3. If $\kappa = 0$, the network activity is one-dimensional, and simply reproduces the pattern of external inputs. If $\kappa \neq 0$, the network response is instead a non-trivial two-dimensional combination of the input and connectivity structure patterns. In general, the value of κ, and therefore the organization of network activity, depends on the geometric arrangement of the input vector I with respect to the connectivity structure vectors m and n, as well as on the strength of the random component of the connectivity g.

A non-vanishing κ, together with non-trivial two-dimensional activity, can be obtained from a variety of configurations of the vectors I, m and n. To start with, we consider the geometrical configuration obtained when the structure vectors m and n are orthogonal to each other (Fig. 7.3 **b**). In that case, the overlap between them is zero, and the spontaneous activity in the network bears no sign of the underlying connectivity structure. Adding an external input can however reveal this structure and generate non-trivial two-dimensional activity. This happens if the input vector I has a non-zero overlap with the left-structure vector n. In such a situation, the activity ϕ of the network will have a component along n because of the inputs, leading to a non-zero overlap κ, which from Eq. 7.4 implies that the network activity will have a component along the right-structure vector m. Increasing the external input along the direction of n will therefore increase the output along m (although m and n are orthogonal) (Fig. 7.3 **d**, top), while increasing the input along a direction orthogonal to n will decrease activity along m (Fig. 7.3 **d**, bottom) and even possibly totally eliminate it. Note that irrespective of its direction, an external input tends to suppress chaos present for strong random connectivity (Fig. 7.3 **c**), but whether this suppression is accompanied by an increase in the two-dimensional structure in the activity depends on the direction of the input with respect to the left-structure vector n.

A different geometrical configuration is obtained when the structure vectors m and n have a non-zero overlap along a common direction. As already shown in Fig. 7.2, an overlap larger than unity between m and n will induce a non-zero overlap κ, and non-trivial, structured spontaneous dynamics. Adding an external input will modify κ, but also the nature of the dynamics. Here we focus on the region of parameter space where the strength of disorder g is larger than unity, so that in absence of inputs the network can display structured static activity ($\kappa \neq 0$), structured chaotic activity ($\kappa \neq 0$) or homogeneous chaotic activity ($\kappa = 0$), depending on the strength of the structured connectivity. Adding an external input along the direction of the left-structure vector n progressively suppresses both bistability and fluctuating, chaotic activity (Fig. 7.3 **f**), and amplifies the structure in network activity by increasing the overlap κ (Fig. 7.3 **g**). Large external inputs along the parallel direction therefore reliably set the network into a state in which the activity is a two-dimensional combination of the input direction and the structure direction m. Note that this occurs even if the connectivity structure strength is too weak to induce structured spontaneous activity ($m^T n/N < 1$): adding the external input along the structured direction unveils the structured connectivity and leads to non-zero κ and therefore two dimensional network activity even in that case (Fig. 7.3 **g**, bottom). An external input added along a direction orthogonal to both m and n also suppresses chaotic and bistable activity, but in contrast decreases the overlap κ and the amount of two-dimensional structure.

When the structure vectors m and n overlap, but are not identical, the left-structure vector

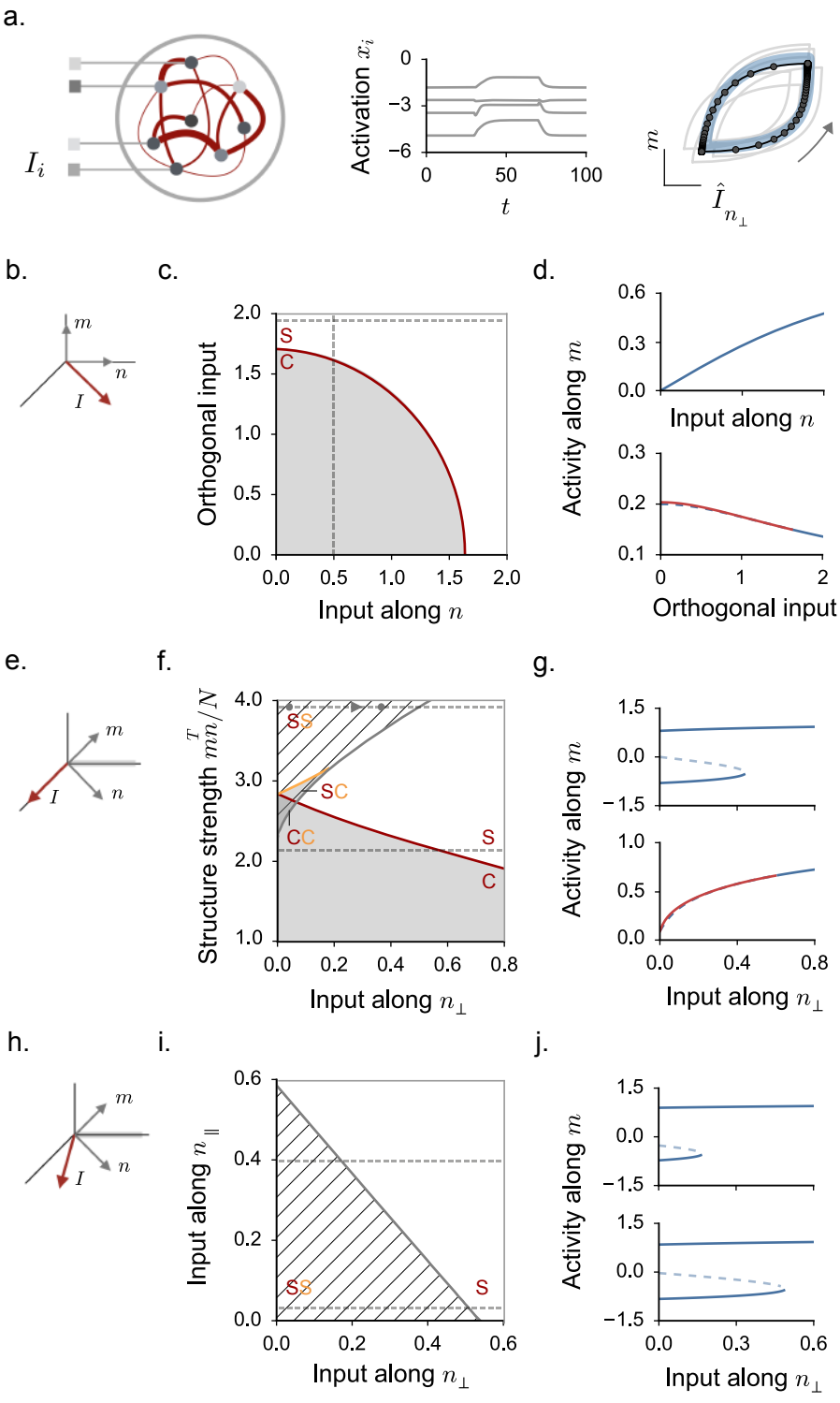

a.
I_i
Activation x_i
0
−3
−6
0
50
100
t
m
$\hat{I}_{n_\perp}$
b.
m
n
I
c.
Orthogonal input
Input along n
S
C
d.
Activity along m
Input along n
Orthogonal input
e.
m
n
I
f.
Structure strength $\frac{T}{mn}/N$
SS
SC
CC
S
C
Input along $n_\perp$
g.
Activity along m
Input along $n_\perp$
h.
m
n
I
i.
Input along $n_\parallel$
SS
S
Input along $n_\perp$
j.
Activity along m
Input along $n_\perp$

Figure 7.3 *(previous page)*: External inputs lead to two-dimensional activity in random networks with unit rank structure. **a.** The pattern of external inputs can be represented by an N-dimensional vector $I = \{I_i\}$, where I_i is the input to unit i. DMF theory shows that the network activity lies on average in the plane defined by the input vector I and the right-structure vector m. The precise organization and the regime of network activity are determined by the geometrical arrangement of the vector I and the structure vectors m and n. Three different cases are illustrated: **b-d.** The structure vectors m and n are orthogonal to each other, and the external input pattern I has a component along n and a component orthogonal to both m and n. **e-g.** The structure vectors m and n have a non-zero overlap. The external input pattern I overlaps with n along its non-shared component (indicated by $n_\perp$), and is thus orthogonal to m. **h-j.** The structure vectors m and n have a non-zero overlap. The external input pattern I is a sum of a component along $n_\parallel$, the direction common to m and n, and a component along $n_\perp$, the direction of n perpendicular to $n_\parallel$. The component along $n_\parallel$ can play the role of a fixed, modulatory input, while the component along $n_\perp$ represents a variable stimulus. **b**, **e**, **h**: summaries of the geometrical arrangement of the three vectors of interest m, n and I in each case. In the case of non-vanishing structure strengths, vectors m and n overlap on the direction corresponding to the shaded axis. **c**, **f**, **i**: phase diagrams showing the type of dynamical activity as function of input and connectivity structure strength. As in Fig. 7.1, shaded areas indicate chaotic dynamics; hatched areas indicate that two stable DMF solutions exist and network activity is bistable. When two stable solutions exist, the yellow and the red letter indicate whether each of them is stationary (S) or chaotic (C). Note that stationary and chaotic dynamics can coexist (SC region). **d**, **g**, **j**: component of network activity along the right-structure vector m, as quantified by the overlap κ (see Eq. 7.4). Parameter values correspond to dashed grey lines in the phase diagrams. Details and colors as in Fig. 7.2. The two rightmost pannels in **a** show transient network dynamics in response to a step input. Left: time traces of the activation variable for four randomly selected units. Right: the population activity is projected onto the plane defined by the vectors m and $\hat{I}_{n_\perp}$, which corresponds to the input direction parallel to $n_\perp$. The start and end parameters are indicated by grey dots in the phase diagram **f**. Light blue trace: theoretical prediction. Grey traces: trajectories of seven different network realizations, $N = 4000$. Individual trajectories deviate from the theoretical prediction because of finite-size fluctuations. The average across different realizations is shown as the black trace. The black points indicate the velocity of the trajectory, as they are equally spaced in time. As in Fig. 7.1, the structure vectors m and n are generated from a Gaussian distribution, and for the sake of simplicity, we consider overlap directions which are aligned with the unitary vector u. In this figure, $g = 2.2$, $\Sigma_m = 1.0$, $\Sigma_n = 1.0$.

n can be decomposed in a sum of two orthogonal components: a component along the overlap ($n_{\parallel}$) and a component perpendicular to the overlap ($n_{\perp}$). The network can therefore receive a superposition of two orthogonal inputs, one along each of these directions (Fig. 7.3 **h**). One of these inputs can for instance represent a stimulus, while the other can play the role of a fixed top-down or contextual modulation. In such a setting, the value of the modulatory input directly controls the extent of the bistable range in response to the stimulus (Fig. 7.3 **i-j**). We will show in the following that this simple non-linear phenomenon can play an important computational role.

So far we have described only the equilibrium state attained after applying an external input for a long time. The two-dimensional nature of the dynamics in response to external inputs is however particularly apparent at the level of temporal responses to inputs. Transient, input-driven dynamics can be analyzed within our theoretical framework by linearizing the dynamics around the corresponding equilibrium state. Our theory predicts that a step input generates two-dimensional trajectories in the $m - I$ plane. The predicted trajectories capture well the average dynamics due to structured connectivity (Fig. 7.3 **a**, right). For a fixed, finite-size network, the directions defined by m and I correspond to the two dominant dimensions of the activity that would be obtained for instance using a dimension-reduction analysis such as Principal Components Analysis [41]. The random part of connectivity leads to additional fluctuations in the remaining $N - 2$ directions.

In summary, external inputs in general suppress chaotic and bistable dynamics (Fig. 7.3 **c**, **f**, **i**), and therefore always decrease the amount of variability in the dynamics. The specific effects of inputs on the structure of network activity however depend on the geometrical arrangement of the pattern of inputs with respect to the connectivity structure vectors m and n. These two structure vectors appear to play different roles. The vector m determines the output pattern of network activity, while the vector n instead selects the inputs that give rise to patterned outputs. An output structured along m can be obtained either when n selects recurrent inputs (non-zero overlap between n and m) or when it selects external inputs (non-zero overlap between n and I).

7.3 The mean field framework

We introduce here the theoretical framework that has been used to derive the results shown in Section 7.1 and 7.2. We first present in detail the model that we adopt, and we later derive the mean field equations that have been solved in order to characterize the network activity.

7.3.1 The network model

We studied the dynamics of a large network of rate units. Similarly to the models adopted in Parts I and II, every unit in the network is characterized by a continuous variable $x_i(t)$, commonly interpreted as the total input current. More generically, we also refer to $x_i(t)$ as the *activation* variable. The output of each unit is a non-linear function of its inputs modeled as a sigmoidal function $\phi(x)$. In line with previous works [127, 132, 73, 108], we use $\phi(x) = \tanh(x)$. In Appendix C we show that qualitatively similar dynamical regimes appear in network models with more realistic, positively defined activation functions. The transformed variable $\phi(x_i(t))$ is interpreted as the firing rate of unit i, and is also referred to as the *activity* variable.

The time evolution is specified by the following dynamics:

$$\dot{x_i}(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij}\phi(x_j(t)) + I_i. \tag{7.5}$$

We focus on a particular class of connectivity matrices, which can be written as a sum of two terms:

$$J_{ij} = g\chi_{ij} + P_{ij}. \tag{7.6}$$

Similarly to [127], χ_{ij} is a Gaussian all-to-all random matrix, where every element is drawn from a centered normal distribution with variance $1/N$. The parameter g scales the strength of random connections in the network. The second term P_{ij} is a low rank matrix. More precisely, we impose $\text{rank}(P_{ij}) \ll N$.

To begin with, we focused on the simplest case where P_{ij} is a rank one matrix, which can generally be written as the external product between two one-dimensional vectors m and n:

$$P_{ij} = \frac{m_i n_j}{N}. \tag{7.7}$$

The theoretical framework is extended to the case of rank two structures in Chapter 8. The entries of vectors m and n are independent of the random bulk of the connectivity χ_{ij}. Note that the only non-zero eigenvalue of P is given by the scalar product $m^T n/N$, and the corresponding right and left eigenvectors are, respectively, vectors m and n. In the following, we will therefore refer to $m^T n/N$ as the structure strength, and to m and n as the right- and left-structure vectors.

As stated in Eq. 7.7, we consider here structured perturbations which scale weakly in the large N limit, i.e. as $1/N$. In contrast, the elements of the all-to-all, random connectivity component χ_{ij} scale as $1/\sqrt{N}$. We will show that such a choice nevertheless results in finite $O(1)$ perturbations of network dynamics.

7.3.2 Computing the network statistics

In this study, we consider the low rank part of the connectivity fixed, while the random part varies between different realizations of the connectivity. The resulting network activity is therefore partially random and partially determined by the structure vectors m and n. Our aim is to characterize the dynamics that emerge from the interplay between these two components as the main parameters of the architecture, the structure strength $m^T n/N$ and the disorder strength g, are varied. We start by examining the activity in absence of external inputs ($I_i = 0$ $\forall i$ in Eq. 7.5).

Our mathematical analysis of network dynamics is based on a Dynamical Mean Field (DMF) approach [127, 99, 69, 58] (see also Chapter 2), which allows us to derive an effective description of the activity states by averaging over the disorder originating from the random part of the connectivity. Across different realizations of the random connectivity matrix χ_{ij}, the network dynamics are characterized in terms of a probability distribution, whose first- and second-order statistics are computed self-consistently.

The DMF theory relies on the hypothesis that a disordered component in the coupling structure, here represented by χ_{ij}, efficiently decorrelates single neuron activity when the network is sufficiently large [16, 90]. In this limit, each unit obeys a Langevin-like equation:

$$\dot{x_i}(t) = -x_i(t) + \eta_i(t), \tag{7.8}$$

where the forcing term η_i is given by a Gaussian process. This Gaussian process can in principle have different first and second-order statistics for each unit, but is otherwise independently drawn across different units. We will show that the hypothesis of decorrelated activity is self-consistent for the specific network architecture we study.

Within the DMF theory, each variable $x_i(t)$ is therefore the solution of a time-dependent random process, and is thus fully determined by the statistics of the effective noise $\eta_i(t)$. In our framework, from Eq. 7.5, we have:

$$\eta_i(t) = g\sum_{j=1}^{N}\chi_{ij}\phi(x_j(t)) + \frac{m_i}{N}\sum_{j=1}^{N} n_j\phi(x_j(t)). \tag{7.9}$$

As in standard DMF derivations, we characterize self-consistently the distribution of η_i by averaging over different realizations of the random matrix χ_{ij} [127, 99, 89]. In the following, $[.]$ indicates an average over the realizations of the random matrix χ_{ij}, while $\langle . \rangle$ stands for an average over different units of the network. Note that the network activity can be equivalently characterized in terms of input current variables $x_i(t)$ or their non-linear transforms $\phi(x_i(t))$. As these two quantities are not independent, the statistics of the distribution of the latter can be written in terms of the statistics of the former.

The mean of the effective noise received by unit i is given by:

$$[\eta_i(t)] = g\sum_{j=1}^{N}[\chi_{ij}\phi(x_j(t))] + \frac{m_i}{N}\sum_{j=1}^{N} n_j[\phi(x_j(t))]. \tag{7.10}$$

Under the hypothesis that in large networks, neural activity decorrelates (more specifically, that activity $\phi(x_j(t))$ is independent of its outgoing weights), we have:

$$[\eta_i(t)] = g\sum_{j=1}^{N}[\chi_{ij}][\phi(x_j(t))] + \frac{m_i}{N}\sum_{j=1}^{N} n_j[\phi(x_j(t))] = m_i\kappa \tag{7.11}$$

as $[\chi_{ij}] = 0$. Here we introduced

$$\kappa \coloneqq \frac{1}{N}\sum_{j=1}^{N} n_j[\phi(x_j(t))] = \langle n_j[\phi_j(t)]\rangle, \tag{7.12}$$

which quantifies the overlap between the mean population activity vector and the left structure vector n. In the last equation, we adopted the short-hand notation $\phi_i(t) \coloneqq \phi(x_i(t))$.

Similarly, the noise correlation function is given by

$$\begin{aligned}[\eta_i(t)\eta_j(t+\tau)] =& g^2\sum_{k=1}^{N}\sum_{l=1}^{N}[\chi_{ik}\chi_{jl}][\phi(x_k(t))\phi(x_l(t+\tau))] \\ &+ \frac{m_i m_j}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} n_k n_l[\phi(x_k(t))\phi(x_l(t+\tau))].\end{aligned} \tag{7.13}$$

Note that every cross-term in the product vanish since $[\chi_{ij}] = 0$. Similarly to standard DMF derivations [127], the first term on the r.h.s. vanishes for cross-correlations ($i \neq j$) while it

survives in the auto-correlation function ($i = j$), as $[\chi_{ik}\chi_{jl}] = \delta_{ij}\delta_{kl}/N$. We get:

$$[\eta_i(t)\eta_j(t+\tau)] = \delta_{ij}g^2\langle[\phi_i(t)\phi_i(t+\tau)]\rangle + \frac{m_i m_j}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} n_k n_l[\phi(x_k(t))\phi(x_l(t+\tau))]. \tag{7.14}$$

We focus now on the second term in the right-hand side. The corresponding sum contains N terms where $k = l$. This contribution vanishes in the large N limit because of the $1/N^2$ scaling. According to our starting hypothesis, when $k \neq l$, activity decorrelates: $[\phi_k(t)\phi_l(t+\tau)] = [\phi_k(t)][\phi_l(t+\tau)]$. To the leading order in N, we get:

$$\begin{aligned}[\eta_i(t)\eta_j(t+\tau)] &= \delta_{ij}g^2\langle[\phi_i(t)\phi_i(t+\tau)]\rangle + \frac{m_i m_j}{N^2}\sum_{k} n_k[\phi(x_k(t))]\sum_{l\neq k} n_l[\phi(x_l(t+\tau))] \\ &= \delta_{ij}g^2\langle[\phi_i(t)\phi_i(t+\tau)]\rangle + m_i m_j\kappa^2\end{aligned} \tag{7.15}$$

so that:

$$[\eta_i(t)\eta_j(t+\tau)] - [\eta_i(t)][\eta_j(t)] = \delta_{ij}g^2\langle[\phi_i(t)\phi_i(t+\tau)]\rangle. \tag{7.16}$$

We therefore find that the statistics of the effective input are uncorrelated across different units, so that our initial hypothesis is self-consistent.

To conclude, for every unit i, we computed the first- and the second-order statistics of the effective input $\eta_i(t)$. The expressions we obtained show that the individual noise statistics depend on the statistics of the full network activity. In particular, the mean of the effective input depends on the average overlap κ, but varies from unit to unit through the components of the right-structure vector m. On the other hand, the auto-correlation of the effective input is identical for all units, and determined by the population-averaged firing rate auto-correlation $\langle[\phi_i(t)\phi_i(t+\tau)]\rangle$.

Once the statistics of $\eta_i(t)$ have been determined, a self-consistent solution for the activation variable $x_i(t)$ can be derived by solving the Langevin-like stochastic process from Eq. 7.8. As a first step, we look at its stationary solutions, which correspond to the fixed points of the original network dynamics.

7.3.3 Dynamical Mean Field equations for stationary solutions

For any solution that does not depend on time, the mean μ_i and the variance Δ_0^I of the variable x_i with respect to different realizations of the random connectivity coincide with the statistics of the effective noise η_i. From Eqs. 7.11 and 7.16, the mean μ_i and variance Δ_0^I of the input to unit i therefore read

$$\begin{aligned}\mu_i &\coloneqq [x_i] = m_i\kappa \\ \Delta_0^I &\coloneqq [x_i^2] - [x_i]^2 = g^2\langle[\phi_i^2]\rangle\end{aligned} \tag{7.17}$$

while any other cross-variance $[x_i x_j] - [x_i][x_j]$ vanishes. We conclude that, on average, the structured connectivity P_{ij} shapes the network activity along the direction specified by its right eigenvector m. Such a heterogeneous stationary state critically relies on a non-vanishing overlap κ between the left eigenvector n and the average population activity vector $[\phi]$. Across different realizations of the random connectivity, the input currents x_i fluctuate around these

mean values. The typical size of fluctuations is determined by the individual variance Δ_0^I, equal for every unit in the network (Fig. 7.4 **a**).

The r.h.s. of Eq. 7.17 contains two population averaged quantities, the overlap κ and the second moment of the activity $\langle[\phi_i^2]\rangle$. To close the equations, these quantities need to be expressed self-consistently. Averaging Eq. 7.17 over the population, we get expressions for the population-averaged mean μ and variance Δ_0 of the input:

$$\begin{aligned}\mu &\coloneqq \langle[x_i]\rangle = M_m\kappa \\ \Delta_0 &\coloneqq \langle[x_i^2]\rangle - \langle[x_i]\rangle^2 = g^2\langle[\phi_i^2]\rangle + (\langle m_i^2\rangle - \langle m_i\rangle^2)\kappa^2.\end{aligned} \tag{7.18}$$

Note that the total population variance Δ_0 is a sum of two terms: the first term, proportional to the strength of the random part of connectivity, coincides with the individual variability Δ_0^I which emerges from different realizations of χ_{ij}; the second term, proportional to the variance of the right-structure vector m, coincides with the variance induced at the population level by the spread of the mean values $\mu_i \propto m_i$. When the vector m is homogeneous ($m_i = \bar{m}$), input currents x_i are centered around the same mean value μ, and the second variance term vanishes.

We next derive appropriate expression for the r.h.s. terms κ and $\langle[\phi_i^2]\rangle$. To start with, we rewrite $[\phi_i]$ by substituting the average over the random connectivity with the equivalent Gaussian integral:

$$\begin{aligned}[\phi_i] &= \int dx_i \phi(x_i) \\ &= \int \mathcal{D}z \phi(\mu_i + \sqrt{\Delta_0^I} z)\end{aligned} \tag{7.19}$$

where we used the short-hand notation $\int \mathcal{D}z = \int_{-\infty}^{+\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \mathrm{d}\, z$. To obtain κ, $[\phi_i]$ needs to be multiplied by n_i and averaged over the population. This average can be expressed by representing the fixed vectors m and n through the joint distribution of their elements over the components:

$$p(m,n) = \frac{1}{N}\sum_{j=1}^{N} \delta(m - m_j)\delta(n - n_j). \tag{7.20}$$

This leads to

$$\begin{aligned}\kappa &= \langle n_i \int \mathcal{D}z \phi(\mu_i + \sqrt{\Delta_0^I} z)\rangle \\ &= \int \mathrm{d}\, m \int \mathrm{d}\, n\, p(m,n)\, n \int \mathcal{D}z \phi(m\kappa + \sqrt{\Delta_0^I} z).\end{aligned} \tag{7.21}$$

Similarly, a suitable expression for the second-order momentum of the firing rate is given by:

$$\langle[\phi_i^2]\rangle = \int \mathrm{d}\, m\, p(m) \int \mathcal{D}z \phi^2(m\kappa + \sqrt{\Delta_0^I} z). \tag{7.22}$$

Eqs. 7.21 and 7.22, combined with Eq. 7.18, provide a closed set of equations for determining κ and Δ_0 once the vectors m and n have been specified.

To further simplify the problem, we reduce the full distribution $p(m,n)$ of elements m_i and n_i to their first- and second-order momenta. That is equivalent to substituting the probability density $p(m,n)$ with a bivariate Gaussian distribution. We therefore write:

$$\begin{aligned}m &= M_m + \Sigma_m\sqrt{1-\rho}\, x_1 + \Sigma_m\sqrt{\rho}\, y \\ n &= M_n + \Sigma_n\sqrt{1-\rho}\, x_2 + \Sigma_n\sqrt{\rho}\, y\end{aligned} \tag{7.23}$$

where x_1, x_2 and y are three normal Gaussian processes. Here, M_m (resp. M_n) and Σ_m (resp. Σ_n) correspond to the mean and the standard deviation of m (resp. n), while the covariance between m and n is given by $\langle m_i n_i \rangle - M_m M_n = \Sigma_m \Sigma_n \rho$. Within a geometrical interpretation, M_m and M_n are the projections of $N-$dimensional vectors m and n onto the unitary vector $u = (1, 1, ...1)/N$, $\Sigma_m \sqrt{\rho}$ and $\Sigma_n \sqrt{\rho}$ are the projection onto a direction direction orthogonal to u and common to m and n, and $\Sigma_m \sqrt{1-\rho}$ and $\Sigma_n \sqrt{1-\rho}$ scale the parts of m and n that are mutually orthogonal (Fig. 7.4 **a**).

The expression for κ becomes:

$$\begin{aligned} \kappa = & \int \mathcal{D}y \int \mathcal{D}x_2 \, (M_n + \Sigma_n \sqrt{1-\rho} x_2 + \Sigma_n \sqrt{\rho} y) \\ & \times \int \mathcal{D}z \int \mathcal{D}x_1 \, \phi(\kappa(M_m + \Sigma_m \sqrt{1-\rho} x_1 + \Sigma_m \sqrt{\rho} y) + \sqrt{\Delta_0^I} z) \end{aligned} \tag{7.24}$$

which gives rise to three terms when expanding the sum $M_n + \Sigma_n \sqrt{1-\rho} x_2 + \Sigma_n \sqrt{\rho} y$. The first term can be rewritten as:

$$\begin{aligned} & M_n \int \mathcal{D}z \, \phi(M_m \kappa + \sqrt{\Delta_0^I + \Sigma_m^2 \kappa^2} z) \\ & = M_n \int \mathcal{D}z \, \phi(\mu + \sqrt{\Delta_0} z) \\ & = M_n \langle [\phi_i] \rangle, \end{aligned} \tag{7.25}$$

which coincides with the overlap between vectors n and $[\phi]$ along the unitary direction $u = (1, 1, ...1)/N$. In the last step, we rewrote our expression for κ in terms of the population averaged statistics μ and Δ_0 (Eq. 7.18).

The second term vanishes, while the third one gives:

$$\begin{aligned} & \Sigma_n \sqrt{\rho} \int \mathcal{D}y \, y \int \mathcal{D}z \int \mathcal{D}x_1 \, \phi(\kappa(M_m + \Sigma_m \sqrt{1-\rho} x_1 + \Sigma_m \sqrt{\rho} y) + \sqrt{\Delta_0^I} z) \\ & = \kappa \rho \Sigma_m \Sigma_n \langle [\phi_i'] \rangle \end{aligned} \tag{7.26}$$

which coincides with the overlap between n and $[\phi]$ in a direction orthogonal to u. Here we used the equality:

$$\int \mathcal{D}z \, z f(z) = \int \mathcal{D}z \, \frac{\mathrm{d} f(z)}{\mathrm{d} z} \tag{7.27}$$

which is obtained by integrating by parts.

Through a similar reasoning we obtain:

$$\langle [\phi_i^2] \rangle = \int \mathcal{D}z \, \phi^2(\mu + \sqrt{\Delta_0} z) \tag{7.28}$$

as in standard DMF derivations.

To conclude, the mean field description of stationary solutions reduces to the system of three implicit equations for μ, κ and Δ_0:

$$\begin{aligned} \mu &= M_m \kappa \\ \Delta_0 &= g^2 \langle [x_i^2] \rangle + \Sigma_m^2 \kappa^2 \\ \kappa &= M_m \langle [\phi_i] \rangle + \kappa \rho \Sigma_m \Sigma_n \langle [\phi_i'] \rangle. \end{aligned} \tag{7.29}$$

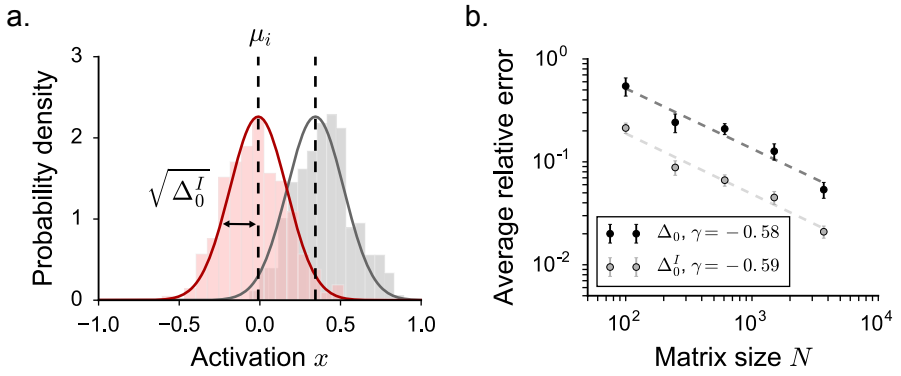


Figure 7.4: Dynamical Mean Field description of stationary solutions. **a.** Across different realizations of the random connectivity χ_{ij}, the activation variables x_i fluctuate around their mean values μ_i. The typical size of such deviations is given by the individual variance Δ_0^I. Continuous lines: distributions obtained for two sample units by solving the DMF equations once the structure term P_{ij} has been fixed. Histograms: numerical distribution from finite-size networks with the same structure P_{ij}. $N_{tr} = 300$ different realizations of the random connectivity term χ_{ij} have been simulated. **b.** Mismatch between the statistics measured in finite-size networks (x_{sim}) and the theoretical prediction (x_{th}) as the network size N is increased. The error is normalized: $|x_{sim} - x_{th}|/x_{th}$. For every realization of P_{ij}, Δ_0^I is measured across 100 different realizations of the random bulk χ_{ij}. Average over 10 rank one structures. The error bars (as in every other figure, if not differently specified) correspond to the standard deviation of the mean. Dashed lines: power-law best fit ($y \propto N^\gamma$). The values of γ are indicated in the legend. Choice of the parameters: $g = 0.6$, $\rho = 0$, $M_m M_n = 2.2$, $\Sigma_m = 2.0$, $\Sigma_n = 1.0$.

Both averages $\langle [.] \rangle$ are performed with respect to a Gaussian distribution of mean μ and variance Δ_0. Once μ, Δ_0 and κ have been determined, the single unit mean μ_i and the individual variance Δ_0^I are obtained from Eq. 7.17.

The dynamical mean field equations given in Eq. 7.29 can be fully solved to determine stationary solutions. Detailed descriptions of these solutions are provided further down for two particular cases: (i) overlap between m and n only along the unitary direction u ($M_m \neq 0$, $M_n \neq 0$, $\rho = 0$); (ii) overlap between m and n only in a direction orthogonal to u ($M_m = M_n = 0$, $\rho \neq 0$). Here we just note that in general, comparisons with simulations shows that the DMF values that we obtain by solving the system in 7.29 approximate well the statistics of finite-size networks. The mismatch between the two decreases in average as the size of the network N is increased (Fig. 7.4 **b**). Note that for unit rank structures, although we used a Gaussian approximation for m and n when computing the averages, our calculation gives good results also when the distribution of m and n is strongly non-Gaussian (see Appendix E).

7.3.4 Transient dynamics and stability of stationary solutions

We now turn to transient dynamics around fixed points, and to the related problem of evaluating whether the stationary solutions found within DMF are stable with respect to the original network dynamics (Eq. 7.5).

For any given realization of the connectivity matrix, the network we consider is completely

deterministic. We can then study the local, transient dynamics by linearizing the dynamics around any stationary solution. We therefore look at the time evolution of a small displacement away from the fixed point: $x(t) = x_i^0 + x_i^1(t)$. For any generic stationary solution $\{x_i^0\}$ the linearized dynamics are given by the stability matrix S_{ij} which reads:

$$S_{ij} = \phi'(x_j^0) \left(g\chi_{ij} + \frac{m_i n_j}{N} \right). \tag{7.30}$$

If the real part of every eigenvalue of S_{ij} is smaller than unity, the perturbation decays in time and thus the stationary solution is stable.

Homogeneous stationary solutions We first consider homogeneous stationary solutions, for which $x_i^0 = \bar{x}$ for all units. A particular homogeneous solution is the trivial solution $x_i = 0$, which the network admits for all parameter values when the transfer function is $\phi(x) = \tanh(x)$. Other homogeneous solutions can be obtained when the vector m is homogeneous, i.e. $m_i = \bar{m}$ for all i.

For homogeneous solutions, the stability matrix reduces to a scaled version of the connectivity matrix J_{ij}:

$$S_{ij} = \phi'(\bar{x}) J_{ij}. \tag{7.31}$$

We are thus left with the problem of evaluating the eigenspectrum of the global connectivity matrix J_{ij}. The matrix J_{ij} consists of a full-rank component χ_{ij}, the entries of which are drawn at random, and of a structured component of small dimensionality with fixed entries. We focus on the limit of large networks; in that limit, an analytical prediction for the spectrum of its eigenvalues can be derived.

Because of the $1/N$ scaling, the matrix norm of P_{ij} is bounded as N increases. We can then apply results from random matrix theory [135] which predict that, in the large N limit, the eigenspectra of the random and the structured parts don't interact, but sum together. The eigenspectrum of J_{ij} therefore consists of two separated components, inherited respectively from the random and the structured terms (Fig. 7.5 **a**). Similarly to [54, 136], the random term χ_{ij} returns a set of $N-1$ eigenvalues which lie on the complex plane in a compact circular region of radius g. In addition to this component, the eigenspectrum of J_{ij} contains the non-zero eigenvalues of P_{ij}: in the case of a rank one matrix, one single outlier eigenvalue is centered at the position $\sum_i m_i n_i / N = \langle m_i n_i \rangle$. In Fig. 7.5 **b** we measure both the outlier position and the radius of the compact circular component. We show that deviations from the theoretical predictions are in general small and decay to zero as the system size is increased.

Going back to the stability matrix $S_{ij} = \phi'(\bar{x}) J_{ij}$, we conclude that a homogeneous stationary solution can lose stability in two different ways, when either $m^T n/N$ or g become larger than $1/\phi'(\bar{x})$. We expect different kinds of instabilities to occur in the two cases. When g crosses the instability line, a large number of random directions become unstable at the same time. As in [127], this instability is expected to lead to the onset of irregular temporal activity. When the instability is lead by the outlier, instead, the trivial fixed point becomes unstable in one unique direction given by the corresponding eigenvector. When $g = 0$, this eigenvector coincides exactly with m. For finite values of the disorder g, the outlier eigenvector fluctuates depending on the random part of the connectivity, but remains strongly correlated with m (Fig. 7.5 **c**), which therefore determines the average direction of the instability. Above the instability, as the network dynamics is completely symmetric with respect to a change of sign of the input variables, we expect the non-linear boundaries to generate two symmetric stationary solutions.

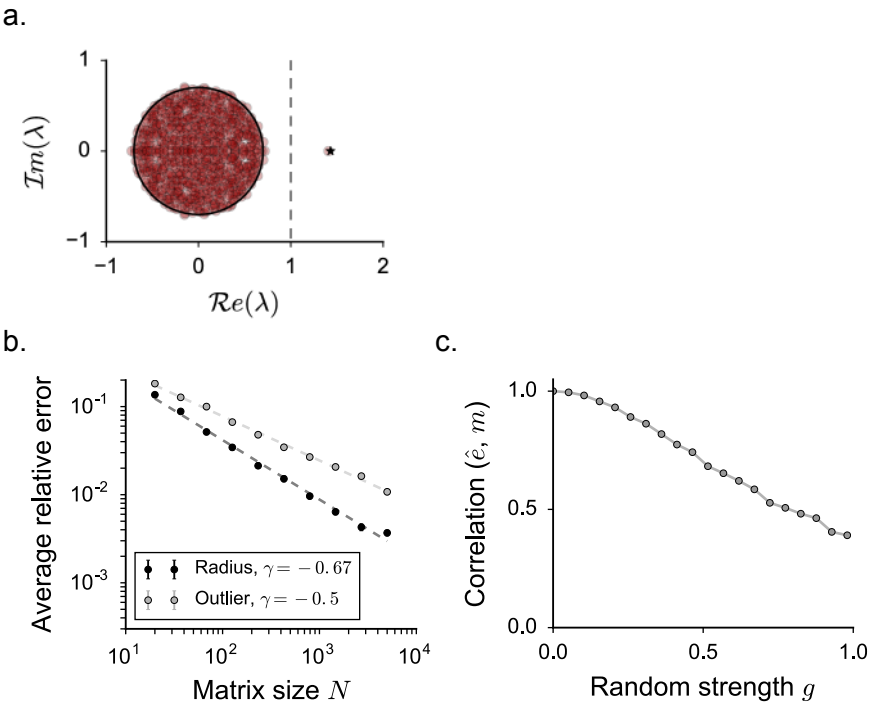


Figure 7.5: Eigenspectrum of the partially structured connectivity matrix J_{ij}, related to the stability matrix S_{ij} of the homogeneous fixed points through: $S_{ij} = \phi'(\bar{x})J_{ij}$. **a.** Eigenspectrum of J_{ij} in the complex plane. Red dots: eigenspectrum of a single realization J_{ij} of size $N = 1000$. In black: theoretical prediction. Every matrix J_{ij} consists of a random and of a fixed structure term. In the large matrix limit, the two eigenspectra sum together. The black circle has radius equal to the total random strength g, and the black star indicates the position of the non-zero eigenvalue of the rank one structure P_{ij}. **b.** Finite size deviations from the theoretical prediction as the matrix size is increased. Details as in Fig. 7.4 **b**. The error is measured across $N_{tr} = 100$ finite size matrices. **c.** Pearson correlation coefficient between the structure eigenvector m and the eigenvector $\hat{e}$ which corresponds to the outlier eigenvalue. Choice of the parameters: $\rho = 0$, $M_m M_n = 1.43$, $\Sigma_m = 0.33$, $\Sigma_n = 0.8$. In **a** and **b**, $g = 0.7$.

Heterogeneous stationary solutions A second type of possible stationary solutions are heterogeneous fixed points, in which different units reach different equilibrium values. For such fixed points, the linearized stability matrix S_{ij} is obtained by multiplying each column of the connectivity matrix J_{ij} by a different gain value (see Eq. 7.30), so that the eigenspectrum of S_{ij} is not identical to the spectrum of J_{ij}.

Numerical investigations reveal that, as for J_{ij}, the eigenspectrum of S_{ij} consists of two discrete components: one compact set of $N-1$ eigenvalues contained in a circle on the complex plane, and a single isolated outlier eigenvalue (Fig. 7.6 **a**).

As previously noticed in [58], the radius of the circular compact set r can be computed as in [98, 6, 5] by summing the variances of the distributions in every column of S_{ij}. To the

leading order in N:

$$r = g\sqrt{\sum_{j=1}^{N} \phi'^2(x_j^0)} \tag{7.32}$$

which, in large networks, can be approximated by the mean field average:

$$r = g\sqrt{\langle [\phi_i'^2] \rangle}. \tag{7.33}$$

Note that, because of the weak scaling in P_{ij}, the structured connectivity term does not appear explicitly in the expression for the radius. As the structured part of the connectivity determines the heterogeneous fixed point, the value of r however depends implicitly on the structured connectivity term through $\langle [\phi_i'^2] \rangle$, which is computed as a Gaussian integral over a distribution with mean μ and variance Δ_0 given by Eq. 7.29. In Fig. 7.6 **a-b** we show that Eq. 7.33 approximates well the radius of finite-size, numerically computed eigenspectra. Whenever the mean field theory predicts instabilities led by r, we expect the network dynamics to converge to irregular non-stationary solutions. Consistently, at the critical point, where $r = 1$, the DMF equations predict the onset of temporally fluctuating solutions (see later on in Section 7.3).

We now turn to the problem of evaluating the position of the outlier eigenvalue. In the case of heterogeneous fixed points, the structured and the random components of the matrix S_{ij} are strongly correlated, as they both scale with the multiplicative factor $\phi'(x_j^0)$, which correlates with the particular realization of the random part of the connectivity χ_{ij}. As a consequence, χ_{ij} cannot be considered as a truly random matrix with respect to $m_i\phi(x_j^0)n_j/N$, and in contrast to the case of homogeneous fixed points, results from [135] do not hold.

We determined numerically the position of the outlier in finite-size eigenspectra (Fig. 7.6 **a-d**). We found that its value indeed significantly deviates from the only non-zero eigenvalue of the rank one structure $m_i\phi(x_j^0)n_j/N$, which can be computed in the mean field framework (when $\rho = 0$, it corresponds to $M_m M_n \langle [\phi_i'] \rangle + M_n \kappa \Sigma_m^2 \langle [\phi_i''] \rangle$). On the other hand, the value of the outlier coincides exactly with the eigenvalue of $m_i\phi(x_j^0)n_j/N$ whenever the random component χ_{ij} is shuffled (black dots in Fig. 7.6 **d**). This observation confirms that the position of the outlier critically depends on the correlations existing between the rank one structure $m_i\phi(x_j^0)n_j/N$ and its specific realization of the random bulk χ_{ij}.

Mean field analysis of transient dynamics and stability of stationary solutions As for heterogeneous fixed points we were not able to assess the position of the outlying eigenvalue using random matrix theory, we turned to a mean field analysis to determine transient activity. This analysis allowed us to determine accurately the position of the outlier, and therefore the stability of heterogeneous fixed points. The approach exploited here is based on [69].

We consider the stability of the single units activation variable x_i when averaged across different realizations of the random connectivity and its random eigenmodes. Directly averaging the network dynamics defined in Eq. 7.5 yields the time evolution of the mean activation μ_i of unit i:

$$\dot{\mu_i}(t) = -\mu_i(t) + m_i\kappa(t). \tag{7.34}$$

We observe that we can write: $\mu_i(t) = m_i\tilde{\kappa}(t)$, where $\tilde{\kappa}$ is the low-pass filtered version of κ: $(1 + \mathrm{d}/\mathrm{d}\,t)\tilde{\kappa}(t) = \kappa(t)$. Small perturbations around the fixed point solution read: $\mu_i(t) = \mu_i^0 + \mu_i^1(t)$. The equilibrium values μ_i^0 correspond to the DMF stationary solution computed

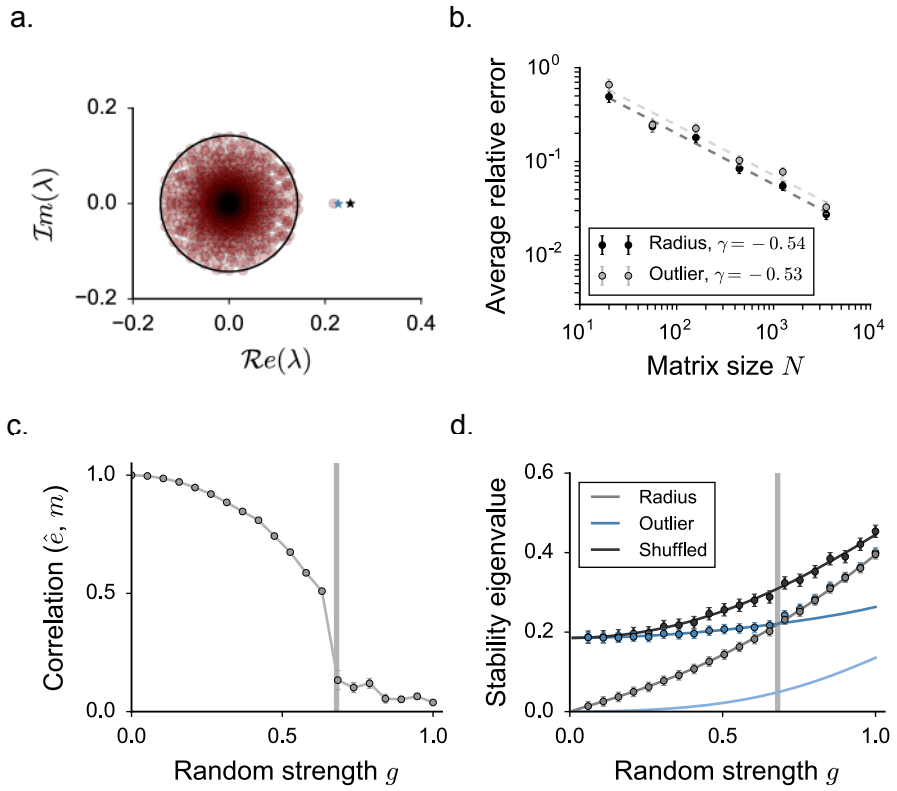


Figure 7.6: Analysis of the eigenspectrum of the linear stability matrix S_{ij} for heterogeneous stationary solutions. **a.** Eigenspectrum of S_{ij} in the complex plane. Red dots: eigenspectrum of a single, finite-size realization of S_{ij}, $N = 2500$. The radius of the black circle corresponds to the theoretical prediction $r = g\langle[\phi_i'^2]\rangle^{\frac{1}{2}}$. The black star indicates the position of the non-zero eigenvalue of the rank one structure $m_i\phi(x_j^0)n_j/N$, which deviates significantly from the position of the outlier eigenvalue. We thus address the problem of evaluating the position of the outlier eigenvalue through a mean field stability analysis, whose prediction is indicated by the blue star. **b.** Mismatch between the results from simulations and mean field predictions for the radius and the outlier position. The error is measured as an average over $N_{tr} = 30$ finite size matrices, and decays as the system size is increased. Details as in Fig. 7.4 **b**. **c.** Pearson correlation coefficient between the structure eigenvector m and the outlier eigenvector $\hat{e}$ as the random strength g is increased and more disorder is injected into the network. Vertical line: at large g values, the outlier eigenvalue get absorbed by the bulk, so that its position cannot be directly measured. **d.** Radius and outlier of the stability eigenspectrum for increasing random strength values. The dots indicate the results of numerical simulations from $N = 2500$ networks, averaged over $N_{tr} = 30$ trials. In grey: radius of the compact bulk (continuous line: mean field prediction r). In blue: position of the outlier eigenvalue (continuous dark and light lines: first and second eigenvalue of matrix $\mathcal{M}$ given in Eq. 7.60). In black: position of the outlier when χ_{ij} is shuffled (continuous line: mean field prediction for the outlier of the structured part $m_i\phi(x_j^0)n_j/N$). Choice of the parameters: $\rho = 0$, $M_mM_n = 2.2$, $\Sigma_m = 0.4$, $\Sigma_n = 2$. In **a** and **b**, $g = 0.5$.

from Eq. 7.17 and 7.29: $\mu_i^0 = m_i\kappa^0$. The first-order perturbations thus obey:

$$\dot{\mu_i}^1(t) = -\mu_i^1(t) + m_i\kappa^1(t), \tag{7.35}$$

indicating that the decay time scale of the mean activity is inherited by the decay time constant of κ^1. An additional equation for the time evolution of κ^1 thus needs to be derived.

When activity is perturbed, the non-linear transform ϕ_i can be evaluated at the first order: $\phi_i^0 \to \phi_i^0 + \phi_i^1(t) = \phi(x_i^0) + \phi'(x_i^0)x_i^1(t)$. As a consequence, the first-order in κ reads:

$$\kappa^1(t) = \langle n_i[\phi'(x_i^0)x_i^1(t)]\rangle. \tag{7.36}$$

Summing Eq. 7.36 to its time-derivative, we get:

$$\dot{\kappa}^1(t) = -\kappa^1(t) + (1 + \frac{\mathrm{d}}{\mathrm{d}\,t})\langle n_i[\phi'(x_i^0)x_i^1(t)]\rangle. \tag{7.37}$$

In order to simplify the r.h.s., we start by considering the average with respect the random part of the connectivity for a single unit i. In order to compute $[\phi'(x_i^0)x_i^1]$, we explicitly build x_i^0 and $x_i^t \coloneqq x_i(t)$ as Gaussian variables centered respectively in μ_i^0 and μ_i^t, which are proportional to m_i. The individual variances of their distributions are time-dependent (Δ_0^{I0} and Δ_0^{It}), and they share some correlated variability $\Delta^{I,t0} = [x_i^t x_i^0] - [x_i^t][x_i^0]$. As a result, we can write:

$$\begin{aligned} x_i^0 &= \mu_i^0 + \sqrt{\Delta_0^{I0} - \Delta^{I,t0}}x_1 + \sqrt{\Delta^{I,t0}}y \\ x_i^t &= \mu_i^t + \sqrt{\Delta_0^{It} - \Delta^{I,t0}}x_2 + \sqrt{\Delta^{I,t0}}y \end{aligned} \tag{7.38}$$

so that, for the first-order response, we get:

$$x_i^1 = \mu_i^1 + \sqrt{\Delta_0^{It} - \Delta^{I,t0}}x_2 - \sqrt{\Delta_0^{I0} - \Delta^{I,t0}}x_1. \tag{7.39}$$

As in classical DMF derivations [127, 99, 69], x_1, x_2 and y are standard normal variables. By integrating over their distributions we can write:

$$\begin{aligned} [\phi'(x_i^0)x_i^1] = \int \mathcal{D}x_1 \int \mathcal{D}x_2 \left(\mu_i^1 + \sqrt{\Delta_0^{It} - \Delta^{I,t0}}x_2 - \sqrt{\Delta_0^{I0} - \Delta^{I,t0}}x_1\right) \\ \times \int \mathcal{D}y\phi'\left(\mu_i^0 + \sqrt{\Delta_0^{I0} - \Delta^{I,t0}}x_1 + \sqrt{\Delta^{I,t0}}y\right). \end{aligned} \tag{7.40}$$

Integrating by parts as in Eq. 7.27 we finally get:

$$[\phi'(x_i^0)x_i^1] = \mu_i^1[\phi_i'] + \left(\Delta^{I,t0} - \Delta_0^{I0}\right)[\phi_i''] \tag{7.41}$$

where the Gaussian integrals $[\phi_i']$ and $[\phi_i'']$ are evaluated at the fixed point.

Note that, at the fixed point, $\Delta^{I,t0} = \Delta_0^{I0}$. As a consequence, $\Delta^{I,t0} - \Delta_0^{I0}$ gives a first-order response:

$$\Delta^{I,10} \coloneqq \Delta^{I,t0} - \Delta_0^{I0} = [x_i^1 x_i^0] - [x_i^1][x_i^0] = [x_i^1 x_i^0] - \mu_i^0\mu_i^1 \tag{7.42}$$

which can be rewritten as a function of the global second-order statistics $\Delta^{10} = \langle [x_i^1 x_i^0] \rangle - \langle [x_i^1] \rangle \langle [x_i^0] \rangle$ as:

$$\begin{aligned} \Delta^{I,10} &= \Delta^{10} - \{ \langle \mu_i^1 \mu_i^0 \rangle - \langle \mu_i^1 \rangle \langle \mu_i^0 \rangle \} \\ &= \Delta^{10} - \Sigma_m^2 \tilde{\kappa}^0 \tilde{\kappa}^1. \end{aligned} \tag{7.43}$$

Furthermore, we observe that we can write:

$$\Delta^{10} = \frac{1}{2} \Delta_0^1 \tag{7.44}$$

which allows us to rewrite the second-order statistics in terms of equal-time variance perturbations: $\Delta_0^1 = \Delta_0^t - \Delta_0^0$. Eq. 7.44 derives from considering that, by definition:

$$\begin{aligned} \Delta^{10} &= \sum_{j=1}^{N} x_j^1 \frac{\partial \Delta^{t0}}{\partial x_j^t} \Big|_0 \\ \Delta_0^1 &= \sum_{j=1}^{N} x_j^1 \frac{\partial \Delta_0^t}{\partial x_j^t} \Big|_0 \end{aligned} \tag{7.45}$$

and by observing that when the derivatives are evaluated at the fixed point, we have:

$$\frac{\partial \Delta^{t0}}{\partial x_j^t} \Big|_0 = \frac{1}{2} \frac{\partial \Delta_0^t}{\partial x_j^t} \Big|_0. \tag{7.46}$$

Eq. 7.41 thus becomes:

$$[\phi'(x_i^0) x_i^1] = m_i \, \tilde{\kappa}^1 [\phi_i'] + \left(\frac{\Delta_0^1}{2} - \Sigma_m^2 \tilde{\kappa}^0 \tilde{\kappa}^1 \right) [\phi_i'']. \tag{7.47}$$

In a second step, we perform the average across different units of the population, by writing m and n as in Eq. 7.23. After some algebra, we get:

$$\begin{aligned} \langle n_i [\phi'(x_i^0) x_i^1(t)] \rangle = \; & \tilde{\kappa}^1 \left[(M_m M_n + \rho \Sigma_m \Sigma_n) \langle [\phi_i'] \rangle + \rho \kappa^0 M_m \Sigma_m \Sigma_n \langle [\phi_i''] \rangle \right] \\ & + \frac{\Delta_0^1}{2} \left[M_n \langle [\phi_i''] \rangle + \rho \kappa^0 \Sigma_m \Sigma_n \langle [\phi_i'''] \rangle \right] \\ := \; & \tilde{\kappa}^1 a + \Delta_0^1 \, b \end{aligned} \tag{7.48}$$

where constants a and b were defined as:

$$\begin{aligned} a &= (M_m M_n + \rho \Sigma_m \Sigma_n) \langle [\phi_i'] \rangle + \rho \kappa^0 M_m \Sigma_m \Sigma_n \langle [\phi_i''] \rangle \\ b &= \frac{1}{2} \left\{ M_n \langle [\phi_i''] \rangle + \rho \kappa^0 \Sigma_m \Sigma_n \langle [\phi_i'''] \rangle \right\}. \end{aligned} \tag{7.49}$$

The time evolution of κ can be finally rewritten as:

$$\dot{\kappa}^1(t) = -\kappa^1(t) + (1 + \frac{\mathrm{d}}{\mathrm{d}\, t}) \left\{ \tilde{\kappa}^1 a + \Delta_0^1 b \right\}, \tag{7.50}$$

so that the time evolution of the perturbed variance must be considered as well.

In order to isolate the evolution law of Δ_0, we rewrite the activation variable $x_i(t)$ by separating the uniform and the heterogeneous components: $x_i(t) = \mu(t) + \delta x_i(t)$. The time evolution for the residual $\delta x_i(t)$ is given by:

$$\dot{\delta x_i}(t) = -\delta x_i(t) + g \sum_{j=1}^{N} \chi_{ij} \phi(x_j(t)) + (m_i - M_m)\kappa(t) \tag{7.51}$$

so that, squaring:

$$\begin{aligned} &\left(\frac{\mathrm{d}\,\delta x_i(t)}{\mathrm{d}\,t}\right)^2 + 2\delta x_i(t)\frac{\mathrm{d}\,\delta x_i(t)}{\mathrm{d}\,t} + \delta x_i(t)^2 = g^2 \sum_{j=1}^{N}\sum_{k=1}^{N} \chi_{ij}\chi_{ik}\phi(x_j(t))\phi(x_k(t)) \\ &+ (m_i - M_m)^2\kappa(t)^2 + g(m_i - M_m)\kappa(t)\sum_{k=1}^{N}\chi_{ij}\phi(x_k(t)). \end{aligned} \tag{7.52}$$

Averaging over i and the realizations of the disorder yields:

$$\begin{aligned} \frac{\mathrm{d}\,\Delta_0(t)}{\mathrm{d}\,t} &= -\Delta_0(t) + g^2\langle[\phi_i^2(t)]\rangle + \Sigma_m^2\kappa(t)^2 - \langle\left[\left(\frac{\mathrm{d}\,\delta x_i(t)}{\mathrm{d}\,t}\right)^2\right]\rangle \\ &:= -\Delta_0(t) + G(\mu, \Delta_0, \kappa) - \langle\left[\left(\frac{\mathrm{d}\,\delta x_i(t)}{\mathrm{d}\,t}\right)^2\right]\rangle \end{aligned} \tag{7.53}$$

as by definition we have: $\langle[\delta x_i^2(t)]\rangle = \Delta_0(t)$.

Expanding the dynamics of Δ_0 to the first order, we get:

$$\dot{\Delta_0^1}(t) = -\Delta_0^1(t) + \mu^1 \frac{\partial G}{\partial \mu}\Big|_0 + \Delta_0^1 \frac{\partial G}{\partial \Delta_0}\Big|_0 + \kappa^1 \frac{\partial G}{\partial \kappa}\Big|_0. \tag{7.54}$$

Note that we could neglect the contributions originating from the last term of Eq. 7.53 because they do not enter at the leading order. Indeed we have:

$$\frac{\partial}{\partial \mu}\langle\left[\left(\frac{\mathrm{d}\,\delta x_i(t)}{\mathrm{d}\,t}\right)^2\right]\rangle\Big|_0 = 2\langle\left[\frac{\mathrm{d}\,\delta x_i(t)}{\mathrm{d}\,t}\frac{\partial}{\partial \mu}\frac{\mathrm{d}\,\delta x_i(t)}{\mathrm{d}\,t}\right]\rangle\Big|_0 = 0 \tag{7.55}$$

since temporal derivatives for every i vanish when evaluated at the fixed point.

A little algebra returns the last three linear coefficients:

$$\begin{aligned} \frac{\partial G}{\partial \mu}\Big|_0 &= 2g^2\langle[\phi_i\phi_i']\rangle \\ \frac{\partial G}{\partial \Delta_0}\Big|_0 &= g^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} \\ \frac{\partial G}{\partial \kappa}\Big|_0 &= 2\Sigma_m^2\kappa^0. \end{aligned} \tag{7.56}$$

Collecting all the results together in Eq. 7.50 we obtain:

$$\dot{\kappa}^1(t) = -\kappa^1(t) + a\kappa^1(t) + b\left\{\mu^1 \frac{\partial G}{\partial \mu}\Big|_0 + \Delta_0^1 \frac{\partial G}{\partial \Delta_0}\Big|_0 + \kappa^1 \frac{\partial G}{\partial \kappa}\Big|_0\right\}. \tag{7.57}$$

By averaging Eq. 7.34 we furthermore obtain:

$$\dot{\mu^1}(t) = -\mu^1(t) + M_m \kappa^1. \tag{7.58}$$

We finally obtained that the perturbation time scale is determined by the population-averaged dynamics:

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\begin{pmatrix}\mu^1\\ \Delta_0^1\\ \kappa^1\end{pmatrix} = -\begin{pmatrix}\mu^1\\ \Delta_0^1\\ \kappa^1\end{pmatrix} + \mathcal{M}\begin{pmatrix}\mu^1\\ \Delta_0^1\\ \kappa^1\end{pmatrix} \tag{7.59}$$

where the evolution matrix $\mathcal{M}$ is defined as:

$$\mathcal{M} = \begin{pmatrix} 0 & 0 & M_m \\ 2g^2\langle[\phi_i\phi_i']\rangle & g^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} & 2\Sigma_m^2\kappa^0 \\ 2bg^2\langle[\phi_i\phi_i']\rangle & bg^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} & b2\Sigma_m^2\kappa^0 + a \end{pmatrix}. \tag{7.60}$$

Note that one eigenvalue of matrix $\mathcal{M}$, which corresponds to the low-pass filtering between κ and μ, is always fixed to zero.

Eqs. 7.59 and 7.60 reveal that, during the relaxation to equilibrium, the transient dynamics of the first- and second-order statistics of the activity are tighly coupled. Diagonalising $\mathcal{M}$ allows to retrieve the largest decay timescale of the network, which indicates the average, structural stability of stationary states.

When an outlier eigenvalue is present in the eigenspectrum of the stability matrix S_{ij}, the largest decay time scale from $\mathcal{M}$ predicts its position. The corresponding eigenvector $\hat{e}$ contains indeed a structured component along m, which is not washed out by averaging across different realizations of χ_{ij}.

The second non-zero eigenvalue of $\mathcal{M}$, which vanishes at $g = 0$, measures a second and smaller effective timescale, which derives from averaging across the remaining $N - 1$ random modes.

Varying g, we computed the largest eigenvalue of $\mathcal{M}$ for corresponding stationary solutions of mean field equations. In Fig. 7.6 **d** we show that, when the stability eigenspectrum includes an outlier eigenvalue, its position is correctly predicted by the largest eigenvalue of $\mathcal{M}$. The mismatch between the two values is small and can be understood as a finite-size effect (Fig. 7.6 **b**, grey).

To conclude, we found that the stability of arbitrary stationary solutions can be assessed by evaluating, with the help of mean field theory, both the values of the radius (Eq. 7.33) and the outlier (Eq. 7.60) of the stability eigenspectrum. Instabilities led by the two different components are expected to reshape activity into two qualitatively different classes of dynamical regimes, which are discussed in detail, further in the chapter, for two specific classes of structures.

7.3.5 Dynamical Mean Field equations for chaotic solutions

When a stationary state loses stability due to the compact component of the stability eigenspectrum, the network activity starts developing irregular temporal fluctuations (for more details, see Chapter 2). Such temporally fluctuating state can be described within the DMF theory by taking into account the full temporal auto-correlation function of the effective noise η_i [127]. For the sake of simplicity, here we derive directly the mean field equations for population-averaged statistics, and we eventually link them back to single unit quantities.

By differentiating twice Eq. 7.8, and by substituting the appropriate expression for the statistics of the noise η_i, we derive that the auto-correlation function $\Delta(\tau) = \langle [x_i(t+\tau)x_i(t)] \rangle - \langle [x_i(t)] \rangle^2$ obeys the second-order differential equation:

$$\ddot{\Delta}(\tau) = \Delta(\tau) - g^2 \langle [\phi_i(t)\phi_i(t+\tau)] \rangle - \Sigma_m^2 \kappa^2. \tag{7.61}$$

In this context, the activation variance Δ_0 coincides with the peak of the full auto-correlation function: $\Delta_0 = \Delta(\tau = 0)$. We expect the total variance to include a temporal term, coinciding with the amplitude of chaotic fluctuations, and a quenched one, representing the spread across the population due to the disorder in χ_{ij} and the structure imposed by the right-structure vector m.

In order to compute the full rate auto-correlation function $\langle [\phi_i(t)\phi_i(t+\tau)] \rangle$, we need to explicitly build two correlated Gaussian variables $x(t)$ and $x(+\tau)$, such that:

$$\begin{aligned} &\langle [x_i(t)] \rangle = \langle [x_i(t+\tau)] \rangle = \mu \\ &\langle [x_i^2(t)] \rangle - \langle [x_i(t)] \rangle^2 = \langle [x_i^2(t+\tau)] \rangle - \langle [x_i(t)] \rangle^2 = \Delta_0 \\ &\langle [x_i(t+\tau)x_i(t)] \rangle - \langle [x_i(t)] \rangle^2 = \Delta(\tau). \end{aligned} \tag{7.62}$$

Following previous studies [127, 99], we obtain:

$$\langle [\phi_i(t)\phi_i(t+\tau)] \rangle = \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta} x + \sqrt{\Delta} z) \right]^2 \tag{7.63}$$

where we used the short-hand notation $\Delta := \Delta(\tau)$ and we assumed for simplicity $\Delta > 0$. As we show later, this requirement is satisfied by our final solution.

In order to visualize the dynamics of the solutions of Eq. 7.61, we study the equivalent problem of a classical particle moving in a one-dimensional potential [127, 99]:

$$\ddot{\Delta}(\tau) = -\frac{\partial V}{\partial \Delta} \tag{7.64}$$

where the potential V is given by an integration over Δ:

$$V(\Delta, \Delta_0) = -\frac{\Delta^2}{2} + g^2 \langle [\Phi_i(t)\Phi_i(t+\tau)] \rangle + \Sigma_m^2 \kappa^2 \Delta \tag{7.65}$$

and $\Phi(x) = \int_{-\infty}^{x} \phi(x')\, \mathrm{d}\, x'$. As the potential V depends self-consistently on the initial condition Δ_0, the shape of the auto-correlation function $\Delta(\tau)$ depends parametrically on the value of Δ_0. Similarly to previous works, we isolate the solutions that decay monotonically from Δ_0 to an asymptotic value $\Delta(\tau \to \infty) := \Delta_\infty$, where Δ_∞ is determined by $\mathrm{d}\, V / \mathrm{d}\, \Delta|_{\Delta = \Delta_\infty} = 0$. This translates into the first condition to be imposed. A second equation comes from the energy conservation condition: $V(\Delta_0, \Delta_0) = V(\Delta_\infty, \Delta_0)$. Combined with the usual equation for the mean μ and the overlap κ, the system of equations to be solved becomes:

$$\begin{aligned} &\mu = M_m \kappa \\ &\kappa = M_n \langle [\phi_i] \rangle + \rho \kappa \langle [\phi_i'] \rangle \\ &\frac{\Delta_0^2 - \Delta_\infty^2}{2} = g^2 \left\{ \int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0} z) - \int \mathcal{D}z \left[\int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 \right\} \\ &+ \Sigma_m^2 \kappa^2 (\Delta_0 - \Delta_\infty) \\ &\Delta_\infty = g^2 \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 + \Sigma_m^2 \kappa^2. \end{aligned} \tag{7.66}$$

The temporally fluctuating state is therefore described by a closed set of equations of the mean activity μ, the overlap κ, the zero-lag variance Δ_0 and the long-time variance Δ_∞. The difference between $\Delta_0 - \Delta_\infty$ represents the amplitude of temporal fluctuations. If temporal fluctuations are absent, $\Delta_0 = \Delta_\infty$, and the system of equations we just derived reduces to the DMF description for stationary solutions given in Eq. 7.29.

A similar set of equations can be derived for single unit activity. As for static stationary states, the mean activity of unit i is given by

$$\mu_i = m_i \kappa. \tag{7.67}$$

The static variance around this mean activity is identical for all units and given by

$$\Delta_\infty^I = g^2 \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 = \Delta_\infty - \Sigma_m^2 \kappa^2 \tag{7.68}$$

while the temporal component Δ_T^I of the variance is identical to the population averaged temporal variance

$$\Delta_T^I = \Delta_0 - \Delta_\infty. \tag{7.69}$$

To conclude, similarly to static stationary states, the structured connectivity P_{ij} shapes network activity in the direction defined by its right eigenvector m whenever the overlap κ does not vanish. For this reason, the mean field theory predicts in some parameter regions the existence of more than one chaotic solution (for further details, see Appendix D). A formal analysis of the stability properties of the different solutions has not been performed. We nevertheless observe from numerical simulations that chaotic solutions tend to inherit the stability properties of the stationary solution they develop from. Specifically, when an homogeneous solution generates two heterogeneous bistable ones, we notice that the former loses stability in favour of the latter.

We finally observe that the critical coupling at which the DMF theory predicts the onset of chaotic fluctuations can be computed by imposing that, at the critical point, the concavity of the potential function $V(\Delta)$ is inverted [127, 58]:

$$\frac{\mathrm{d}^2 V(\Delta, \Delta_0)}{\mathrm{d}\,\Delta^2}\Big|_{\Delta_\infty} = 0 \tag{7.70}$$

and the temporal component of the variance vanishes: $\Delta_0 = \Delta_\infty$. These two conditions are equivalent to the expression: $1 = g^2 \langle [\phi_i'^2] \rangle$ where, as we saw, $g^2 \langle [\phi_i'^2] \rangle$ coincides with the value of the radius of the compact component of the stability eigenspectrum (Eq. 7.33). In the phase diagram of Fig. 7.1 **a**, we solved this equation for g to derive the position of the instability boundary from stationary to chaotic regimes (red line).

7.3.6 Structures overlapping on the unitary direction

In this section, we analyze in detail a specific case, in which the structure vectors m and n overlap solely along the unitary direction $u = (1, 1, ...1)/N$. Within the statistical description of vector components, in this situation the joint probability density $p(m, n)$ can be replaced by the product two normal distributions (respectively, $\mathcal{N}(M_m, \Sigma_m^2)$ and $\mathcal{N}(M_n, \Sigma_n^2)$). The mean values M_m and M_n represent the projections of m and n on the common direction u, and the overlap between m and n is given by $M_m M_n$. The components m and n are

otherwise independent, the fluctuations representing the remaining parts of m and n that lie along mutually orthogonal directions. In this situation, the expression for κ simplifies to

$$\begin{aligned}\kappa &= \langle n_i[\phi_i]\rangle \\ &= M_n\langle[\phi_i]\rangle\end{aligned} \tag{7.71}$$

so that a non-zero overlap κ can be obtained only if the mean population activity $\langle[\phi_i]\rangle$ is non-zero. Choosing independently drawn m and n vectors thus simplifies the mean field network description. The main qualitative features resulting from the interaction between the structured and the random component of the connectivity can however already be observed, and more easily understood, within this simplified setting.

Stationary solutions The DMF description for stationary solutions reduces to a system of two non-linear equations for the population averaged mean μ and variance Δ_0:

$$\begin{aligned}\mu &= M_m M_n\langle[\phi_i]\rangle \coloneqq F(\mu,\Delta_0) \\ \Delta_0 &= g^2\langle[\phi_i^2]\rangle + \Sigma_m^2 M_n^2\langle[\phi_i]\rangle^2 \coloneqq G(\mu,\Delta_0).\end{aligned} \tag{7.72}$$

The population averages $\langle[\phi_i]\rangle$ and $\langle[\phi_i^2]\rangle$ are computed as Gaussian integrals similarly to Eq. 7.28. Eq. 7.72 can be solved numerically for μ and Δ_0 by iterating the equations up to convergence, which is equivalent to numerically simulating the two-dimensional dynamical system given by

$$\begin{aligned}\dot{\mu}(t) &= -\mu + F(\mu,\Delta_0) \\ \dot{\Delta}_0(t) &= -\Delta_0 + G(\mu,\Delta_0),\end{aligned} \tag{7.73}$$

since the fixed points of this dynamical system correspond to solutions of Eq. 7.72.

As the system of equations in 7.72 is two-dimensional, we can investigate the number and the nature of stationary solutions through a simple graphical approach (Fig. 7.7). We plot on the $\mu-\Delta_0$ plane the loci of points where the two individual equations

$$\begin{aligned}\mu &= F(\mu,\Delta_0) \\ \Delta_0 &= G(\mu,\Delta_0)\end{aligned} \tag{7.74}$$

are satisfied. In analogy with dynamical systems approaches, we refer to the two corresponding curves as the DMF *nullclines*. The solutions of Eq. 7.72 are then given by the intersections of the two nullclines.

To begin with, we focus on the nullcline defined by the first equation (also referred to as the μ nullcline). With respect to μ, $F(\mu,\Delta_0)$ is an odd sigmoidal function whose maximal slope depends on the value of Δ_0 and M_mM_n. When $g=0$ and $\Sigma_m=0$, the input variance Δ_0 vanishes. In this case, the points of the μ nullcline trivially reduce to the roots of the equation: $\mu = M_mM_n\phi(\mu)$, which admits either one ($M_mM_n<1$), or three solutions ($M_mM_n>1$). Non-zero values of g and Σ_m imply finite and positive values of Δ_0. As Δ_0 increases, the solutions to the equation $\mu = M_mM_n\langle[\phi_i]\rangle$ vary smoothly, delining the full nullcline in the $\mu-\Delta_0$ plane. As in the case without disorder ($g=0$ and $\Sigma_m=0$), for low structure strengths ($M_mM_n<1$), the μ nullcline consists of a unique branch: $\mu=0\ \forall\Delta_0$ (Fig. 7.7 **b**). At high structure strengths ($M_mM_n>1$), instead, its shape smoothly transform into a symmetric pitchfork (Fig. 7.7 **c**).

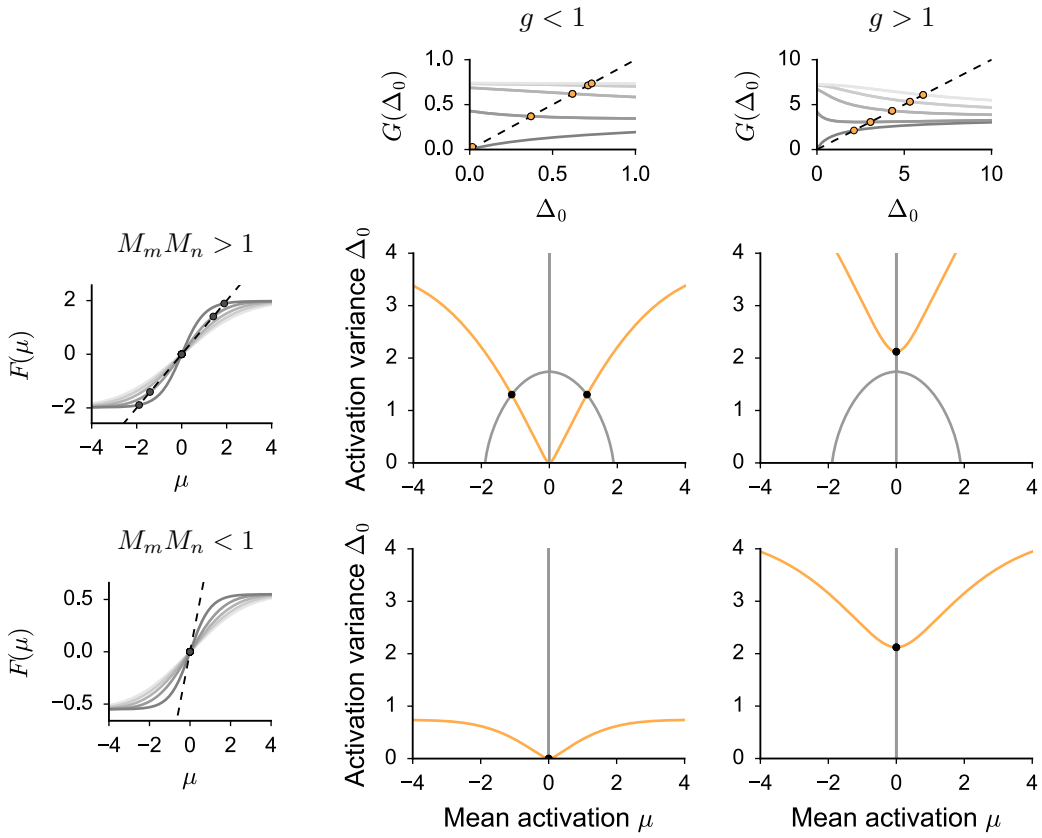


Figure 7.7: Dynamical Mean Field description for partially structured networks whose right and left vectors overlap solely on the unitary direction ($\rho = 0$). Graphical analysis of stationary solutions. Large figures: nullcline plots for the population-averaged DMF equations in 7.72. Black dots indicate the solutions that are stable with respect to the outlier eigenvalue. Four set of parameters (two values for $M_m M_n$, two for g) have been selected. Note that the shape of the μ and the Δ_0 nullcline depends only, respectively, on the value of the structure and the random strengths $M_m M_n$ and g. For the figures in the first (resp. second) row, the structure strength $M_m M_n = 0.55$ (resp. $M_m M_n = 2.0$) is weak (resp. strong). For the figures in the first (resp. second) column: the random strength $g = 0.7$ (resp. $g = 2.0$) is weak (resp. strong). The small figures side associated to every row and column show how the μ (for the rows) and Δ_0 (for the columns) nullclines have been built. We solve $\mu = F(\mu)$ (resp. $\Delta_0 = G(\Delta_0)$) for different initial values of Δ_0 (resp. μ). Different initial conditions are displayed in gray scale. Dark grey refers to $\Delta_0 = 0$ (resp. $\mu = 0$). The dots indicate the solutions for different initial values, which together generate the nullcline curves.

The Δ_0 nullcline is given by the solutions of $\Delta_0 = G(\mu, \Delta_0)$ for Δ_0 as function of μ. As $G(\mu, \Delta_0)$ depends quadratically on μ, the Δ_0 nullcline has a symmetric V-shape centered in $\mu = 0$. The ordinate of its vertex is controlled by the parameter g, as the second term of the second equation in 7.72 vanishes at $\mu = 0$. For $\mu = 0$, the slope of $G(\mu, \Delta_0)$ in $\Delta_0 = 0$ is equal to g^2. As a consequence, for $g < 1$, the vertex of the Δ_0 nullcline is fixed in $(0, 0)$, while for $g > 1$, the vertex is located at $\Delta_0 > 0$ and an isolated point remains at $(0, 0)$.

The stationary solutions of the DMF equations are determined by the intersections be-

tween the two nullclines. For all values of the parameters, the nullclines intersect in $\mu = 0$, $\Delta_0 = 0$, corresponding to the trivial, homogeneous stationary solution. The existence of other solutions are determined by the qualitative features of the individual nullclines, that depend on whether $M_m M_n$ and g are smaller or greater than one (Fig. 7.7). The following qualitative situations can be distinguished: (i) for $M_m M_n < 1$ and $g < 1$, only the trivial solutions exist; (ii) for $M_m M_n > 1$, two additional, symmetric solutions exist for non-zero values of μ and Δ_0, corresponding to symmetric, heterogeneous stationary states; (iii) for $g > 1$, an additional solution exist for $\mu = 0$ and $\Delta_0 > 0$, corresponding to a heterogeneous solution in which individual units have non-zero stationary activity, but the population-average vanishes. For $M_m M_n > 1$, this solution can co-exist with the symmetric heterogeneous ones, but in the limit of large g these solutions disappear (Fig. 7.7).

The next step is to assess the stability of the various solutions. As explained earlier on, the stability of the trivial state $\mu = 0$, $\Delta_0 = 0$ can be readily assessed using random matrix theory arguments (Fig. 7.5). This state is stable only for $M_m M_n < 1$ and $g < 1$. At $M_m M_n = 1$, it loses stability due to the outlying eigenvalue of the stability matrix, leading to the bifurcation already observed at the level of nullclines. At $g = 1$, the instability is due to the radius of the bulk of the spectrum. This leads to a chaotic state, not predicted from the nullclines for the stationary solutions.

The stability of heterogeneous stationary states is assessed by determining separately the radius of the bulk of the spectrum and the position of the outlier (Fig. 7.6). The radius is determined from Eq. 7.33. The outlier is instead computed as the leading eigenvalue of the stability matrix given in Eq. 7.60. Note that, in the present framework, it is possible to show that the latter is equivalent to computing the leading stability eigenvalue of the effective dynamical system introduced in Eq. 7.73, linearized around the corresponding fixed point. The bifurcation obtained when the outlier crosses unity is equivalent to the bifurcation predicted from the nullclines when the symmetric solutions disappear in favor of the heterogeneous solution of mean zero (Fig. 7.7). For $M_m M_n > 1$, we however find that as g is increased, the radius of the bulk of the spectrum always leads to a chaotic instability before the outlier becomes unstable. Correspondingly, the $\mu = 0$ and $\Delta_0 > 0$ stationary state that exist for large g is never stable.

Chaotic solutions For large g, the instabilities of the stationary points generated by the bulk of the spectrum are expected to give rise to chaotic dynamics. We therefore turn to the DMF theory for chaotic states, which are described by an addition variable that quantifies temporal fluctuations. For the case studied here of structure vectors m and n overlapping only along the unitary direction, Eqs. 7.66 become

$$\begin{aligned}
\mu &= F(\mu, \Delta_0, \Delta_\infty) = M_m M_n \int \mathcal{D}z \phi(\mu + \sqrt{\Delta_0} z) \\
\Delta_0 &= G(\mu, \Delta_0, \Delta_\infty) = \Bigg[\Delta_\infty^2 + 2g^2 \Bigg\{ \int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0} z) \\
&- \int \mathcal{D}z \left[\int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 \Bigg\} + M_n^2 \Sigma_m^2 \langle [\phi_i] \rangle^2 (\Delta_0 - \Delta_\infty) \Bigg]^{\frac{1}{2}} \\
\Delta_\infty &= H(\mu, \Delta_0, \Delta_\infty) = g^2 \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 + M_n^2 \Sigma_m^2 \langle [\phi_i] \rangle^2 .
\end{aligned} \tag{7.75}$$

As the system to be solved is now three-dimensional, graphical approaches have only limited use. As for the stationary state, a practical and stable way to find numerically the solutions is to iterate the dynamical system given by

$$\begin{aligned}\dot{\mu} &= -\mu + F(\mu, \Delta_0)\\ \dot{\Delta}_0 &= -\Delta_0 + G(\mu, \Delta_0)\\ \dot{\Delta}_\infty &= -\Delta_\infty + H(\mu, \Delta_\infty).\end{aligned} \tag{7.76}$$

Note that stationary states simply correspond to solutions for which $\Delta_0 = \Delta_\infty$.

As for stationary solutions, different types of chaotic solutions appear depending on the values of the structure strength $M_m M_n$ and the disorder strength g. If $g > 1$ and $M_m M_n < 1$, a single chaotic state exists corresponding to $\mu = 0$ and $\Delta_\infty = 0$, meaning that the temporally averaged activity of all units vanishes, so that fluctuations are only temporal (Fig. 7.1 **b** red). As $M_m M_n$ crosses unity, two symmetric states appear with non-zero values of μ and Δ_∞. These states correspond to bistable heterogeneous chaotic states (Fig. 7.1 **b** orange) that are analogous to bistable heterogeneous stationary states.

The critical disorder strength g_B at which heterogeneous chaotic states emerge (grey boundary in the phase diagram of Fig. 7.1) is computed by evaluating the linear stability of the dynamics in 7.76 around the central solution $(0, \Delta_0, 0)$. A long but straightforward algebra reveals that the stability matrix, evaluated in $(0, \Delta_0, 0)$, is simply given by

$$\begin{pmatrix} M_m M_n \langle \phi' \rangle & 0 & 0 \\ 0 & \frac{g^2\left(\langle \phi^2 \rangle + \langle \Phi \phi' \rangle - \langle \Phi \rangle \langle \phi' \rangle\right)}{\Delta_0} & 0 \\ 0 & 0 & g^2 \langle \phi' \rangle^2 \end{pmatrix}, \tag{7.77}$$

such that g_B corresponds to the value of the random strength g for which the largest of its three eigenvalues crosses unity.

7.3.7 Structures overlapping on an arbitrary direction

In the previous paragraph, we focused on the simplified scenario where the structure vectors m and n overlapped only in the unitary direction. Here, we briefly turn to the opposite case where the overlap along the unitary direction u vanishes (i.e. $M_m = 0$, $M_n = 0$), but the overlap ρ along a direction orthogonal to u is non-zero (Fig. 7.8 **a**). As we will show, although the equations describing the network activity present some formal differences, they lead to qualitatively similar regimes. The same qualitative results apply as well to the general case, where an overlap exists on both the unitary and an orthogonal direction.

The network dynamics can be studied by solving the DMF equations 7.29 and 7.66 by setting $\mu = 0$. Stationary solutions are now determined by:

$$\begin{aligned}\kappa &= \rho \kappa \Sigma_m \Sigma_n \langle [\phi_i'(0, \Delta_0)] \rangle \coloneqq F(\kappa, \Delta_0)\\ \Delta_0 &= g^2 \langle [\phi_i^2(0, \Delta_0)] \rangle + \Sigma_m^2 \kappa^2 \coloneqq G(\kappa, \Delta_0).\end{aligned} \tag{7.78}$$

Note that, in this more general case, the relevant first-order statistics of network activity is given by the overlap κ, which now can take non-zero values even when the population-averaged activity $\langle [\phi_i] \rangle$ vanishes.

As in the previous case, the stationary solutions can be analyzed in terms of nullclines. The main difference lies in the κ nullcline given by $\kappa = \rho \kappa \Sigma_m \Sigma_n \langle [\phi_i'(0, \Delta_0)] \rangle$. As both sides

of the first equation are linear and homogeneous in κ, two classes of solutions exist: a trivial solution ($\kappa = 0$ for any Δ_0), and a non-trivial one ($\Delta_0 = \tilde{\Delta}_0$ for any κ), with $\tilde{\Delta}_0$ determined by:

$$\langle [\phi_i'(0, \tilde{\Delta}_0)] \rangle = 1/(\rho \Sigma_m \Sigma_n). \tag{7.79}$$

Because $0 < \phi'(x) < 1$, Eq. 7.79 admits non-trivial solutions only for sufficiently large overlap values: $\rho > 1/\Sigma_m\Sigma_n$. In consequence, the κ nullcline takes qualitatively different shapes depending on the value of ρ: (i) for $\rho < 1/\Sigma_m\Sigma_n$, it consists only of vertical branch $\kappa = 0$ (ii) for $\rho > 1/\Sigma_m\Sigma_n$ and additional horizontal branch $\Delta_0 = \tilde{\Delta}_0$ appears (Fig. 7.8).

The Δ_0 branch is qualitatively similar to the previously studied case of m and n overlapping along the unitary direction, with a qualitative change when the disorder parameter g crosses unity.

The stationary solutions are given by the intersections between the two nullclines. Although the shape of the κ nullcline is distinct from the shape of the μ nullcline studied in the previous case, qualitatively similar regimes are found. The trivial stationary state $\kappa = 0$, $\Delta_0 = 0$ exists for all parameter values. When the structure strength $\rho\Sigma_m\Sigma_n$ exceeds unity, two symmetric heterogeneous states appear with non-zero κ values of opposite signs (but vanishing mean μ). Finally for large g an additional state appears with $\kappa = 0$, $\Delta_0 > 0$.

Similarly to Fig. 7.2, the solutions of Eq. 7.78, which correspond to stationary activity states, are shown in blue in Fig. 7.9. In Fig. 7.9 **a** we address their stability properties: again we find that when non-centered stationary solutions exist, the central fixed point becomes instable. The instability is led by the outlier eigenvalue of its stability eigenspectrum. Similarly to Fig. 7.1, furthermore, the DMF theory predicts an instability to chaotic phases for high g values. As for stationary states, both heterogeneous and homogeneous chaotic solutions are admitted (Fig. 7.9 **b-c**); heterogeneous chaotic states exist in a parameter region where the values of g and ρ are comparable.

7.3.8 Response to external inputs

To conclude, following Section 7.2, we examine the effect of non-vanishing external inputs on the network dynamics. We consider the situation in which every unit receives a potentially different input I_i, so that the pattern of inputs at the network level is characterized by the N-dimensional vector $I = \{I_i\}$. The network dynamics in general depend on the geometrical arrangement of the vector I with respect to the structure vectors m and n. Within the statistical description used in DMF theory, the input pattern is therefore characterized by the first- and second-order statistics M_I and Σ_I of its elements, as well as by the value of the correlations Σ_{mI} and Σ_{nI} with the vectors m and n. In geometric terms, M_I quantifies the component of I along the unit direction u, while Σ_{mI} and Σ_{nI} quantify the overlaps with m and n along directions orthogonal to u. For the sake of simplicity, we consider two structure vectors m and n that overlap solely on the unitary direction ($\rho = 0$). The two vectors thus read (see Eq. 7.23):

$$\begin{aligned} m &= M_m + \Sigma_m x_1 \\ n &= M_n + \Sigma_n x_2. \end{aligned} \tag{7.80}$$

The input pattern can overlap with the structure vectors on the common (u) and on the orthogonal directions (x_1 and x_2). It can moreover include further orthogonal components of

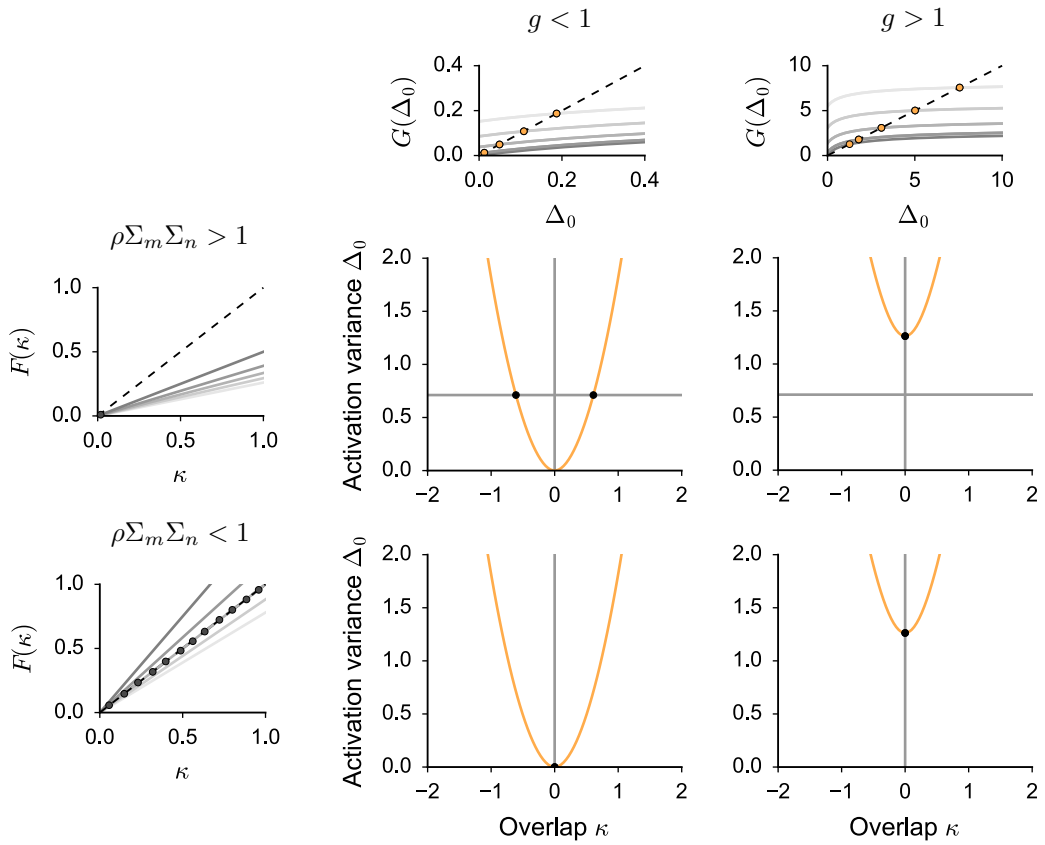


Figure 7.8: Dynamical Mean Field description of stationary solutions for partially structured networks whose right and left vectors overlap onto an arbitrary direction y ($M_m = M_n = 0$). Graphical analysis of stationary solutions. Large figures: nullcline plots for the population-averaged DMF equations in 7.78. Black dots indicate the solutions that are stable with respect to the outlier eigenvalue. Four set of parameters (two values for $\rho\Sigma_m\Sigma_n$, two for g) have been selected. Note that the shape of the κ and the Δ_0 nullcline depends only, respectively, on the value of the structure and the random strengths $\rho\Sigma_m\Sigma_n$ and g. For the figures in the first (resp. second) row, the structure strength $\rho\Sigma_m\Sigma_n$ (resp. $\rho\Sigma_m\Sigma_n$) is weak (resp. strong). For the figures in the first (resp. second) column: the random strength $g = 0.5$ (resp. $g = 1.7$) is weak (resp. strong). The small figures associated to every row and column show how the κ (for the rows) and Δ_0 (for the columns) nullclines have been built. We solve $\kappa = F(\kappa)$ (resp. $\Delta_0 = G(\Delta_0)$) for different initial values of Δ_0 (resp. κ). Different initial conditions are displayed in gray scale. Dark grey refers to $\Delta_0 = 0$ (resp. $\kappa = 0$). The dots indicate the solutions for different initial values, which together generate the nullcline curves.

strength $\Sigma_\perp$. The most general expression for the input vector can thus be written as:

$$I_i = M_I + \frac{\Sigma_{mI}}{\Sigma_m} x_1 + \frac{\Sigma_{nI}}{\Sigma_n} x_2 + \Sigma_\perp h \tag{7.81}$$

where h is a standard normal vector. We first focus on the equilibrium response to constant inputs, and then turn to transient dynamics.

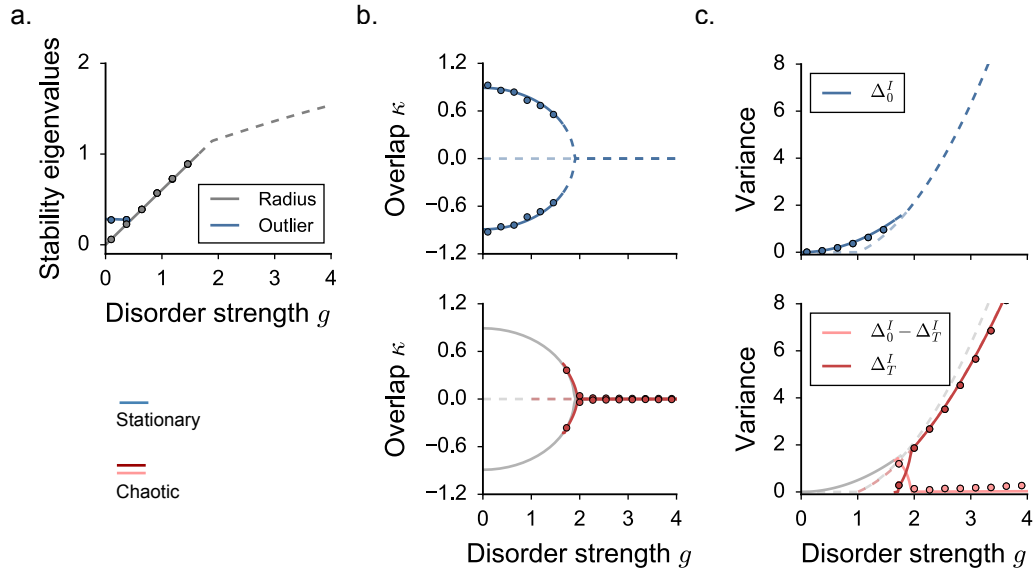


FIGURE 7.9: Dynamical Mean Field description for partially structured networks whose generating vectors m and n overlap onto an arbitrary direction y ($M_m = M_n = 0$). Bifurcation diagram of the network activity statistics as the random strength g is increased. **a.** Stability eigenspectrum of stationary solutions, mean field prediction for the radius of the compact part and the outlier position. **b.** Overlap $\kappa = \langle n_i[\phi_i] \rangle$. **c.** Individual second order statistics. The DMF solutions are displayed as continuous (resp. dashed) lines if they correspond to a stable (resp. unstable) state. In **c-d**, top panels display statistics for stationary solutions and bottom panels display statistics for chaotic solutions. Dots: we measured network activity statistics in finite-size networks, starting from globally positive and negative initial conditions. Activity is integrated up to $T = 600$. $N = 4000$, average over 10 different network realizations. Choice of the parameters: $\Sigma_m = \Sigma_n = 1.5$, $\rho = 2.0/\Sigma_m\Sigma_n$.

The mean field equations in presence of external inputs can be derived in a straightforward fashion by following the same steps as in the input-free case. We start by considering the statistics of the effective coupling term, which is given by $\xi_i(t) = \eta_i(t) + I_i(t)$, with $\eta_i(t)$ defined as in Eq. 7.9. We can then exploit the statistics of $\eta_i(t)$ which have been computed in the previous paragraphs to obtain the equation for the mean activity:

$$\mu_i = [x_i] = m_i\kappa + I_i. \tag{7.82}$$

Eq. 7.82 indicates that the direction of the average network activity is determined by a combination of the structured recurrent connectivity and the external input pattern. The final direction of the activation vector in the N-dimensional population space is controlled by the value of the overlap κ, which depends on the relative orientations of m, n and I. Its value is given by the self-consistent equation:

$$\begin{aligned} \kappa &= \langle n_i[\phi_i] \rangle \\ &= \langle n_i \int \mathcal{D}z \phi(m_i\kappa + I_i + \sqrt{\Delta_0^I} z) \rangle \\ &= M_n \langle [\phi_i] \rangle + \Sigma_{nI} \langle [\phi_i'] \rangle, \end{aligned} \tag{7.83}$$

as both vectors m and I share non-trivial overlap directions with n.

The second-order statistics of the noise are given by:

$$[\xi_i(t)\xi_j(t+\tau)] = \delta_{ij}g^2\langle[\phi_i(t)\phi_i(t+\tau)]\rangle + m_i m_j\kappa^2 + (m_i I_j + m_j I_i)\kappa + I_i I_j. \quad (7.84)$$

Averaging across the population we obtain:

$$\langle[\xi_i(t)\xi_i(t+\tau)]\rangle - \langle[\xi_i(t)]^2 = g^2\langle[\phi_i^2]\rangle + \Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2. \quad (7.85)$$

The first term of the r.h.s. represents the quenched variability inherited from the random connectivity matrix, while $\Sigma_\mu^2 = \Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2$ represents the variance induced by the structure, which is inherited from both vectors m and I (Eq. 7.82). From Eq. 7.81, the variance of the input reads:

$$\Sigma_I^2 = \frac{\Sigma_{mI}^2}{\Sigma_m^2} + \frac{\Sigma_{nI}^2}{\Sigma_n^2} + \Sigma_\perp^2. \quad (7.86)$$

The final DMF equations to be solved are given by the following system:

$$\begin{aligned} \mu &= M_m\kappa + M_I \\ \ddot{\Delta} &= \Delta - \left\{g^2\langle[\phi_i(t)\phi(t+\tau)]\rangle + \Sigma_m^2\kappa^2 + 2\Sigma_{mI}\kappa + \Sigma_I^2\right\} \\ \kappa &= M_n\langle[\phi_i]\rangle + \Sigma_{nI}\langle[\phi_i']\rangle \end{aligned} \quad (7.87)$$

which, similarly to the cases we examined in detail so far, admits both stationary and chaotic solutions. As for spontaneous dynamics, the instabilities to chaos are computed by evaluating the radius of the eigenspectrum of the stability matrix S_{ij} (Eq. 7.33). The stability matrix can admit an outlier eigenvalue as well, whose value can be predicted with a mean field stability analysis. Extending the arguments already presented in the previous paragraphs allows to show that the effective stability matrix $\mathcal{M}$ is given by:

$$\mathcal{M} = \begin{pmatrix} 0 & 0 & M_m \\ 2g^2\langle[\phi_i\phi_i']\rangle & g^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} & 2\Sigma_m^2\kappa^0 + 2\Sigma_{mI} \\ 2bg^2\langle[\phi_i\phi_i']\rangle & bg^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} & b(2\Sigma_m^2\kappa^0 + 2\Sigma_{mI}) + a \end{pmatrix}, \quad (7.88)$$

with:

$$\begin{aligned} a &= M_m M_n\langle[\phi_i']\rangle + M_m\Sigma_{nI}\langle[\phi_i'']\rangle \\ b &= \frac{1}{2}\left\{M_n\langle[\phi_i'']\rangle + \Sigma_{nI}\langle[\phi_i''']\rangle\right\}. \end{aligned} \quad (7.89)$$

As in the input-free case, when the stability eigenspectrum contains one outlier eigenvalue, its position is well predicted by the largest eigenvalue of $\mathcal{M}$.

In the following, we refer to Fig. 7.3 and analyse in detail the contribution of every input direction to the final network dynamics.

In Fig. 7.3 **c**-**d**-**e**, we consider a unit rank structure whose vectors m and n are orthogonal: $M_m = M_n = 0$. The input pattern overlaps with n along x_2, and includes an additional orthogonal component along z ($\Sigma_\perp > 0$). We furthermore assume $\Sigma_{mI} = 0$.

As can be seen from the equation for κ (Eq. 7.87), the overlap between the input and the left vector n has the effect of increasing the value of κ, which would otherwise vanish since the structure has null strength. In response to the input, a structured state emerges. From the same equation, furthermore, one can notice that the Σ_{nI} term has the effect of breaking the

sign reversal symmetry ($x \to -x$) that characterizes the mean field equations in the case of spontaneous dynamics.

On the other hand, increasing the strength of the orthogonal input component $\Sigma_\perp$ does not directly affect the equation for κ. Nevertheless, the orthogonal input $\Sigma_\perp$ tends to increase the value of Δ_0 through Σ_I^2. Since $\langle[\phi'(x)]\rangle$ decreases with Δ_0, larger values of Σ_I imply smaller values of κ. External inputs that are orthogonal to n have thus the effect of reducing structured activity. Note that a similar effect is obtained for external inputs correlating with m along x_1 ($\Sigma_{mI} > 0$).

In the rest of Fig. 7.3, we include non vanishing structure strengths ($M_m, M_n \neq 0$).

In Fig. 7.3 **f**-**g**-**h**, the input pattern overlaps with n on a direction that is orthogonal to the structure overlap ($M_I = 0$, $\Sigma_{nI} > 0$). The external input has in this case three major effects. First, by breaking the sign reversal symmetry, it disrupts the symmetry between the two stable solutions when bistability is created at large structure strengths. Second, it increases the value of Δ_0 through Σ_I, which in turns reduces the extension of the bistability regions in the phase diagram. Third, it tends to suppress chaotic activity.

Finally, external inputs allined with the non-shared (x_2) and the shared (u) directions of n affect the mean field equations in slightly different ways, but effectively influence the dynamics in a very similar fashion. In Fig. 7.3 **i**-**j**-**k**, we include an input component along the structure overlap direction ($M_I > 0$). We show that the contribution coming from positive M_I values sums with the contribution along x_2 given by Σ_{nI}, and contributes to reducing the phase space area corresponding to chaotic and bistable activity. Different input directions along different n components can thus be used to tune the degree of symmetry breaking introduced in the mean field solutions.

Asymmetric solutions A major effect of external inputs is that they break the sign reversal symmetry ($x \to -x$) present in the network dynamics without inputs. As a consequence, in the parameter regions where the network dynamics admit bistable structured states, the two stable solutions are characterized by different statistics and stability properties.

To illustrate this effect, we focus on the simple case where the external input pattern I overlaps with the structure vector m and n solely on the unitary direction ($M_I \neq 0$, $\Sigma_{mI} = \Sigma_{nI} = 0$). The solutions of the system of equations corresponding to stationary states can be visualised with the help of the graphical approach, which unveils the symmetry breaking of network dynamics induced by external inputs (Fig. 7.10).

Similarly to the input free case (Fig. 7.7 and 7.8), the Δ_0 nullcline consists of a symmetric V-shaped curve. In contrast to before, however, the vertex of the nullcline is no longer fixed in $(0, 0)$, but takes positive ordinate values also at low g values. The value of $G(0, \Delta_0)$, indeed, does not vanish, because of the finite contribution from the input pattern Σ_I^2.

The nullcline curves of μ are instead strongly asymmetric. For low $M_m M_n$ values, one single μ nullcline exists. In contrast to the input-free case, this nullcline is no longer centered in zero. As a consequence, it intersects the Δ_0 nullclines in one non-zero point, corresponding to a unique heterogeneous stationary solution. As $M_m M_n$ increases, a second, separated branch can appear. In contrast to the input-free case, the structure strength at which the second branch appears is not always equal to unity, but depends on the mean value of the input. If $M_m M_n$ is strong enough, the negative branch of the nullcline can intersect the Δ_0 nullcline in two different fixed points, while a third solution is built on the positive μ nullcline. As g increases, the two intersections on the negative branch become closer and closer and they

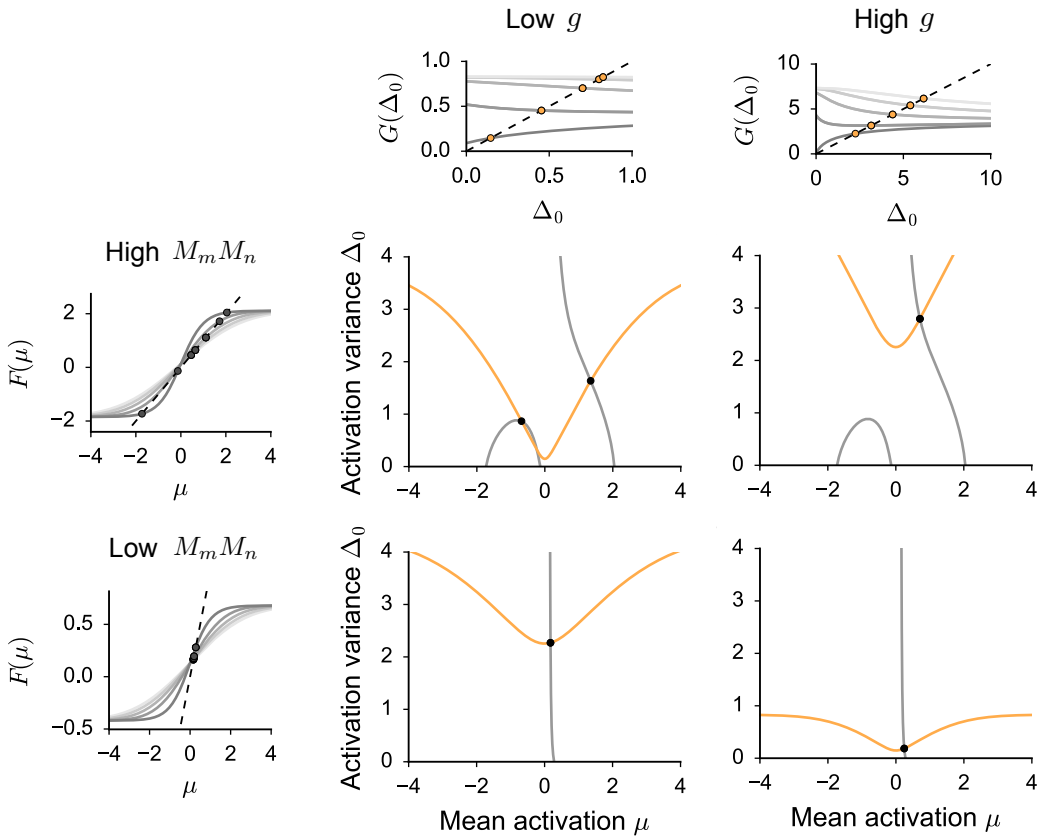


Figure 7.10: Dynamical Mean Field description for partially structured networks whose right and left vectors overlap solely on the unitary direction ($\rho = 0$) in presence of external input patterns. Graphical analysis of stationary solutions. Large figures: nullcline plots for the population-averaged DMF equations in 7.87. Black dots indicate the solutions that are stable with respect to the outlier eigenvalue. Four set of parameters (two values for $M_m M_n$, two for g) have been selected. Note that the shape of the μ and the Δ_0 nullcline depends only, respectively, on the value of the structure and the random strengths $M_m M_n$ and g together with the input statistics. For the figures in the first (resp. second) row, the structure strength $M_m M_n = 0.55$ (resp. $M_m M_n = 2.0$) is weak (resp. strong). For the figures in the first (resp. second) column: the random strength $g = 0.7$ (resp. $g = 2.0$) is weak (resp. strong). The small figures associated to every row and column show how the μ (for the rows) and Δ_0 (for the columns) nullclines have been built. We solve $\mu = F(\mu)$ (resp. $\Delta_0 = G(\Delta_0)$) for different initial values of Δ_0 (resp. μ). Different initial conditions are displayed in gray scale. Dark grey refers to $\Delta_0 = 0$ (resp. $\mu = 0$). The dots indicate the solutions for different initial values, which together generate the nullcline curves. Choice of the parameters: $M_I = 0.13$, $\Sigma_{nI} = \Sigma_{mI} = 0$, $\Sigma_I = 0.3$.

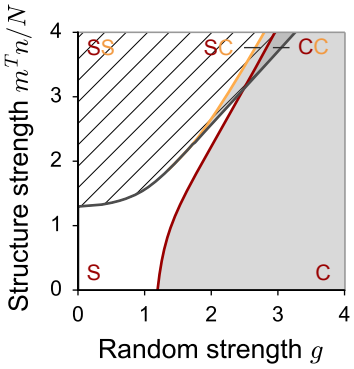


FIGURE 7.11: External inputs disrupt the sign reversal symmetry of the network dynamics: effect on the phase diagram. As in Fig. 7.10, the input vector is aligned with the common direction of the structure vectors m and n, which coincides with the unitary one ($\Sigma_{nI} = \Sigma_{mI} = \Sigma_{\perp} = 0$). Similarly to Fig. 7.1 **a**, we vary the main parameters of the recurrent connectivity: the structure and the random strength. In contrast to Fig. 7.1, however, note that different bistable solutions lose stability on different critical boundaries. As in Fig. 7.1, shaded areas indicate chaotic dynamics; hatched areas indicate that two stable DMF solutions exist and network activity is bistable. When two stable solutions exist, the yellow and the red letter indicate wheter each of them is stationary (S) or chaotic (C). Note that stationary and chaotic dynamics can coexist (SC region). Choice of the parameters: $\Sigma_{nI} = \Sigma_{mI} = 0$, $M_I = 0.1$, $M_m M_n = 2.2$, $\Sigma_m = \Sigma_n = 0$.

eventually collapse together. At a critical value g_B, the network activity discontinuously jumps from negative to positive mean solutions.

As they are no longer symmetrical, the stability of the positive and the negative fixed points has to be assessed separately, and gives raise to different instability boundaries. Computing the position of the outlier reveals that, when more than one solution is admitted by the mean field system of equations, the centered one is always unstable.

As the stability boundaries of different stationary solutions do not necessarily coincide, in presence of external input patterns the phase diagram of the dynamics are in general more complex (Fig. 7.3). Specifically, hybrid dynamical regimes, where one static solution co-exist with a chaotic attractor, can be observed. A phase diagram similar to the one in Fig. 7.1 **a**, which illustrate the dependence on the two main connectivity parameters g and $m^T n/N$, is shown in Fig. 7.11.

Transient dynamics We now turn to transient dynamics evoked by a temporal step in the external input (Fig. 7.3 **b**). We specifically examine the projection of the activation vector and its average onto the two salient directions spanned by vectors m and I.

The transient dynamics of relaxation to a stationary solution can be assessed by linearizing the mean field dynamics. We compute the time course of the average activation vector $\{\mu_i\}$, and we finally project it onto the two orthogonal directions which are indicated in the small insets of Fig. 7.3 **b**.

Similarly to Eq. 7.34, the time evolution of μ_i is governed by:

$$\dot{\mu}_i(t) = -\mu_i(t) + m_i \kappa(t) + I_i(t) \tag{7.90}$$

so that, at every point in time:

$$\mu_i(t) = m_i\tilde{\kappa}(t) + \tilde{I}_i(t), \tag{7.91}$$

where $\tilde{\kappa}(t)$ and $\tilde{I}_i(t)$ coincide with the low-pass filtered versions of $\kappa(t)$ and $I(t)$.

When the network activity is freely decaying back to an equilibrium stationary state, $\tilde{I}_i(t)$ coincides with a simple exponential relaxation to the pattern I_i. The decay time scale is set by the time evolution of activity (Eq. 7.5), which is taken here to be equal to unity:

$$\tilde{I}_i(t) = I_i + (I_i^{ic} - I_i)e^{-t}. \tag{7.92}$$

The time scale of $\tilde{\kappa}(t)$ is inherited by the dynamics of $\kappa(t)$. We thus refer to our mean field stability analysis, and we compute the relaxation time of the population statistics $\kappa(t)$ as the largest eigenvalue of the stability matrix $\mathcal{M}$. The eigenvalue predicts a time constant τ_r, which is in general larger than unity. As a consequence, the relaxation of $\kappa(t)$ obeys, for small displacements:

$$\kappa(t) = \kappa^0 + (\kappa^{ic} - \kappa^0)e^{-\frac{t}{\tau_r}}, \tag{7.93}$$

where the asymptotic value of κ^0 is determined from the equilibrium mean field equations (Eqs. 7.87). Finally, the time course of $\tilde{\kappa}(t)$ is derived as the low-pass filter version of Eq. 7.93 with unit decay time scale.

Implementing computations 8

In Chapter 7, we developed a simple geometric understanding of the dynamics in networks with a given one-dimensional connectivity structure. We now reverse our approach to study how a given computation can be implemented by choosing appropriately the structured part of the connectivity. To this aim, we exploit our understanding of the input-driven dynamics in networks with unit rank structures, and we extend our theoretical framework to specific rank two connectivity setups.

Throughout this chapter, our approach consists of starting by fixing a task, which implies a qualitative input-output relationship to be implemented by the network. We then explicitly design a suitable low-dimensional structure which forces the network to satisfy this rule. This approach results in model networks with low-dimensional dynamics which robustly implement computations, by tolerating large amounts of noise and temporal fluctuations. The computational networks we derive operate in stable dynamical regimes, and our theoretical approach allows to quantitatively determine the expected output and its relaxation time scale.

Note that this approach conceptually differs from the procedures which are typically used in supervised training. Here, the input-output relationship determined by the task is fulfilled only at a qualitative level: in contrast to supervised training, the exact output values are not fixed by the problem. An alternative, training-oriented approach, is adopted and discussed in Chapter 9.

8.1 Computing with unit rank structures: the Go-Nogo task

We consider first the computation underlying one of the most basic behavioral tasks, the Go-Nogo stimulus detection. In this task, an animal has to produce a specific motor response, e.g. press a lever or lick a spout, in response to a specific sensory stimulus (the Go stimulus), and ignore all other stimuli (Nogo stimuli). We will show that a recurrent network with a rank-one connectivity structure provides a simple but computationally powerful implementation of this task. We provide here the main results, and we discuss the details about the theoretical setting in Section 8.1.1.

We model the sensory stimuli as random patterns of external inputs to the network, so that each stimulus is represented by a fixed, randomly-chosen $N-$dimensional vector $I^{(k)}$. To

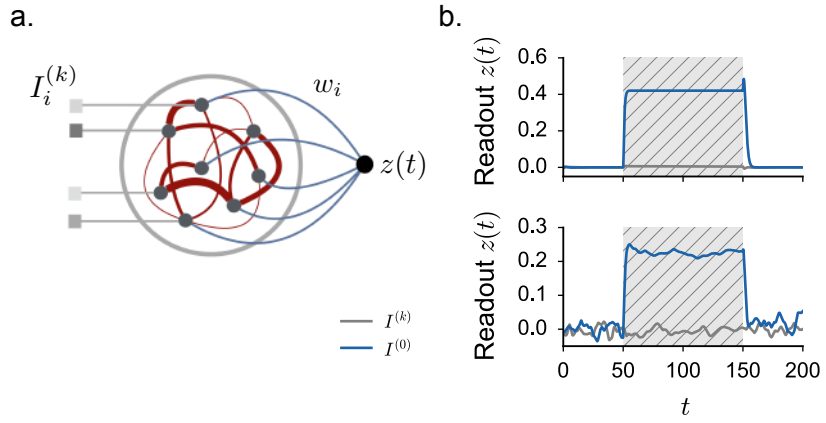


Figure 8.1: Implementing a simple computation with a unit rank connectivity structure: a network model for the Go-Nogo task. **a.** We consider a random network with unit rank connectivity structure. Stimuli are modeled as fixed, randomly generated patterns of inputs $I^{(k)}$. The network output is given by a linear readout $z(t) = \sum_i w_i \phi_i(t)/N$, where the readout weights w_i are fixed and randomly chosen. Our aim is to design a unit rank structure which allows this specific network output to activate selectively in response to the Go pattern $I^{(0)}$. **b.** A possible solution is obtained by selecting $m = w$ and $n = I^{(0)}$. We look at the behavior of a finite-size network ($N = 2500$) in a stationary (top: $g = 0.8$) and in a chaotic (bottom: $g = 2.4$) dynamical regime. In the time window corresponding to the grey hatched region, an input pattern is presented to the network. Blue trace: readout in response to the Go stimulus $I^{(0)}$. Grey trace: readout in response to Nogo inputs $I^{(k)}$ for $k \neq 1$. As shown in Fig. 8.2, most of the network variability is averaged at the level of the readout. In a finite size network, however, small temporal fluctuations can be seen at the level of the readout $z(t)$.

model the motor response, we supplement the network with an output unit, which produces a linear readout $z(t) = \frac{1}{N}\sum_i w_i \phi_i(t)$ of network activity (Fig. 8.1 **a**). The readout weights w_i are chosen randomly and form also a fixed $N-$dimensional vector w. The task of the network is to produce an output that is selective to the Go stimulus: the readout z needs to be non-zero for the input pattern $I^{(0)}$ that corresponds to the Go stimulus, and zero for any other input $I^{(k)}, k > 0$. Moreover, we require that the network output is specific to the chosen readout w, so that reading out network activity along a direction uncorrelated (orthogonal) to w should lead to no output.

Our aim is to determine two N-dimensional vectors m and n that generate the appropriate rank one connectivity structure to implement the task. As shown in Eq. 7.4 and Fig. 7.3, the response of the network to the input pattern $I^{(k)}$ is in general two-dimensional and lies in the plane spanned by the vectors m and $I^{(k)}$. The output unit will produce a non-zero readout only if the readout vector w has a non-vanishing overlap with either m or $I^{(k)}$. As w is assumed to be uncorrelated, and therefore orthogonal, to all input patterns, this implies that the structure vector m needs to have a non-zero overlap with the readout vector w for the network to produce a non-trivial output. This output will depend on the specific input through the overlap κ between the network activity and the left-structure vector n. Assuming this vector is orthogonal to m, as shown in Fig. 7.3, the overlap κ will be non-zero only if n has a non-vanishing overlap with the input pattern $I^{(k)}$. Choosing $m = w$ and $n = I^{(0)}$, we

therefore obtain the simplest rank-one connectivity that implements the desired computation. Such a network generates non-zero activity along the direction m only if the input pattern is $I^{(0)}$, thus implementing selectivity to the Go stimulus (Fig. 8.1 **b**). As the output direction m is aligned with the readout vector w, the readout of the activity along any direction orthogonal to w will be zero, so that the network response is also specific to the fixed readout. Note that within this simple model, the representations are highly distributed over the population, and every unit is by construction selective to a mix of several stimuli and the output [107]. In particular, the network response always contains a component along the direction of the input stimulus, so that information about the input is always present and can be extracted with the appropriate readout. This observation is consistent with the ubiquitous finding that higher cortical areas generally encode both the outcome of a decision and the original stimuli that led to that decision [113, 62].

The determined unit rank connectivity structure implements the scaffold for the desired input-output transform, but the random part of the connectivity adds variability around the target output. As shown in Fig. 7.2 **c**, the fluctuations of the activity of each unit around the value set by the unit rank connectivity structure increase with the strength g of disorder. Summing the activity of individual units through the readout unit however averages out these fluctuations, so that the readout error decreases with network size as $1/\sqrt{N}$ (Fig. 8.2 **a**). For large g, the activity in the network becomes chaotic, but the structured connectivity ensures that the network still performs the required computation, albeit with additional temporal fluctuations in finite-size networks (Fig. 8.1 **b**, bottom).

The simple rank-one implementation of the Go-Nogo discrimination task has very desirable computational properties, in particular in terms of generalization to noisy or novel stimuli. Suppose for instance that the Go stimulus is corrupted with noise, so that the network receives an input pattern that is correlated with the Go stimulus, but not identical to it. With the above choice for the implementation of the task ($m = w$ and $n = I^{(0)}$, w and $I^{(0)}$ uncorrelated), the output of the network to the noisy stimulus will be approximately proportional to the correlation coefficient between the input and the Go pattern (Fig. 8.2 **b**).

A more selective, non-linear readout can be obtained with a slightly different choice of structure vectors, in which m and n still have a non-zero overlap with respectively w and $I^{(0)}$, but also include a non-zero mutual overlap, i.e. a component in a shared direction orthogonal to w and $I^{(0)}$ (Fig. 8.2 **c**). In that case, if the correlation between the input and and the Go stimulus is low, the network will be in a bistable regime (Fig. 7.3 **j-k**) in which the two states will average each other out, so that the readout will be close to zero. In contrast, for inputs strongly correlated with the Go stimulus, only a single state is stable and leads to a strong readout. The output therefore behave in a binary, all-or-none manner, and can implement a finer discrimination between correlated inputs. The threshold that sets the boundary between Go and Nogo responses can in particular be controlled by an additional input. As shown in Fig. 7.3 **i-j**, a fixed input along the direction of the overlap between m and n determines the extent of the bistable region. Changing the value of this input will therefore modulate the position of the threshold, and can even totally suppress the output. This additional input can therefore for instance implement a contextual modulation or gating of the output.

The computation in the network can also be extended in a straightforward way to the detection of a category of Go stimuli, rather than a single stimulus. Suppose two instances of the category are represented by input vectors $I^{(0)}$ and $I^{(1)}$. Choosing the left-structure vector n as the average between $I^{(0)}$ and $I^{(1)}$ will directly implement the selectivity to these two individual stimuli, but also to any intermediate stimulus represented as a linear combination

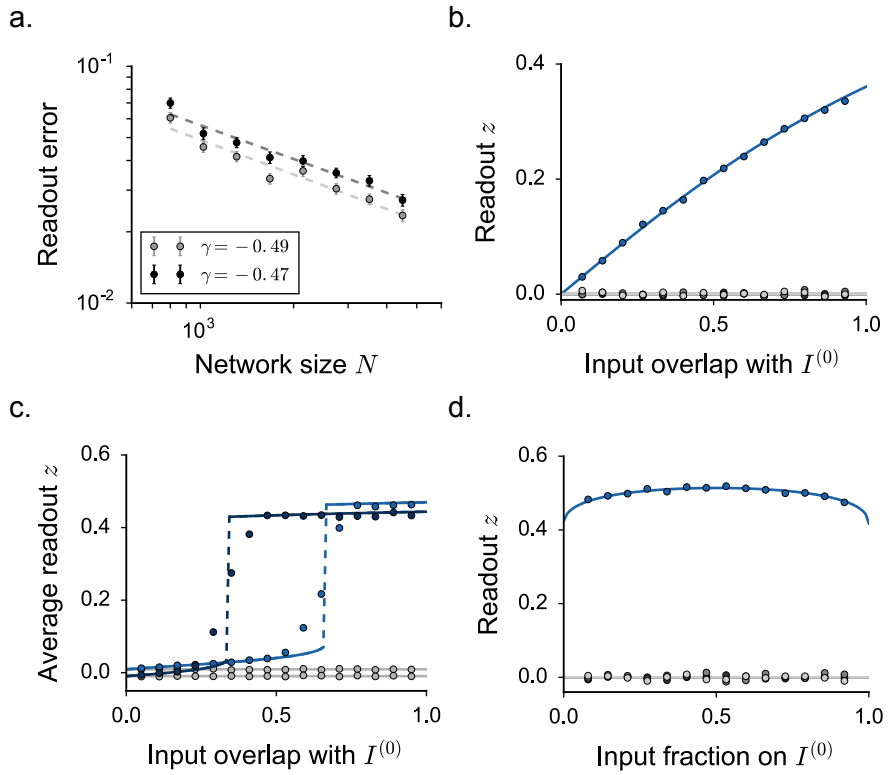


Figure 8.2: Implementing a simple computation with a unit rank connectivity structure: performance of the network model for the Go-Nogo task. **a.** Absolute distance between the theoretical prediction and the value of the readout z obtained from finite-size realizations. As expected, the magnitude of the average normalized error decays with the network size N. In grey: $g = 0.2$, in black: $g = 1$. Dashed lines: power-law best fit ($y \propto N^{\gamma}$). The values of γ are indicated in the legend. **b.** Readout in response to the Go stimulus corrupted by noise. In blue, readout for decoding weights equal to w_i, and a stimulus that includes a non-zero overlap with both the Go input $I^{(0)}$ and an additional noise component. Continuous lines: DMF theoretical prediction; dots: readout value measured from finite-size networks ($N = 3000$, average over 30 different realizations). In grey, readout using a random set of decoding weights. **c.** When the structure vectors m and n share a strong overlap onto a common direction (here $\rho_m = \rho_n = 2.0$, see Section 8.1.1), the readout only responds for overlap values above a given threshold. If the overlap is below threshold, activity is bistable (see Fig. 7.3 **j-k**), and the average is close to zero. We show an average over 60 realizations, where the network activity is initialized from random initial conditions. Dark and light blue correspond to two different values of the modulatory inputs, which have the effect of controlling the position of the transition: $\gamma = \pm 0.2$. **d.** Generalization properties of the selective response. We select two Go stimuli $I^{(0)}$ and $I^{(1)}$, and we set $n = I^{(0)} + I^{(1)}$. We build the input pattern as a normalized mixture of the two preferred patterns, and we gradually increase the component along $I^{(0)}$. The readout $z(t)$ robustly responds to every mixture input, and the small modulation in the output value is masked by finite size fluctuations in finite networks. In this figure, the input and the readout vectors are Gaussian patterns of standard deviation $\Sigma = 1.2$.

of $I^{(0)}$ and $I^{(1)}$. The network therefore automatically generalizes the detection to novel stimuli, that have not been directly represented at the level of connectivity (Fig. 8.2 **d**).

8.1.1 Mean field equations

We provide here a quantitative characterization of the rank one network we have built to perform selective and specific input-output associations as in a Go-Nogo detection task (Fig. 8.1).

In each trial the network receives an input specified by an N-dimensional vector I taken from a set of p possible input vectors $\{I^{(k)}\}_{k=0...p}$, with $p \ll N$. The components of the input patterns are generated independently from a Gaussian distribution of mean zero and variance Σ_I. As the components of the inputs are uncorrelated, the input vectors $\{I^{(k)}\}_{k=0...p}$ are mutually orthogonal in the limit of large N. The network activity is moreover read out linearly through a vector w generated from a Gaussian distribution of mean zero and variance Σ_w, so that the readout value is given by:

$$z = \langle w_i[\phi_i] \rangle. \tag{8.1}$$

Our aim is to determine structure vectors m and n such that: (i) the readout is selective, i.e. $z \neq 0$ if the input is $I^{(0)}$ and $z = 0$ for inputs $I^{(k)}$, $k \geq 1$; (ii) the readout is specific to the vector w, i.e. it is zero for any readout vector uncorrelated with w. We have shown that the simplest network architecture which satisfies these requirements is given by $m = w$ and $n = I^{(0)}$, i.e. the right-structure vector m corresponds to the readout vector, and the left-structure vector corresponds to the preferred stimulus $I^{(0)}$.

The response of the network can be analysed by looking at stationary and chaotic solutions of Eq. 7.87. In the case analyzed here, the structure vectors have no overlap direction, so we set $M_m = M_n = M_I = \Sigma_{mI} = 0$, which implies $\mu = 0$. The first-order network statistics is determined by the overlap Σ_{nI} between the left-structure vector and the input vector. As the left-structure is given by $I^{(0)}$, Σ_{nI} is the overlap between the current input pattern I and the preferred pattern $I^{(0)}$, that we indicate by $\Delta := \langle I_i^{(0)} I_i \rangle$. When varying the amount of noise, the total input variance Σ_I is kept fixed. The total input is therefore constructed as

$$I = I^{(0)} \frac{\Delta}{\Sigma_I^2} + x \sqrt{\Sigma_I^2 - \frac{\Delta^2}{\Sigma_I^4}}, \tag{8.2}$$

where the noise vector x is generated from a normal Gaussian distribution.

From Eq. 7.83 we have:

$$\begin{aligned} \kappa &= \langle n_i[\phi_i] \rangle \\ &= \langle I_i^{(0)}[\phi_i] \rangle \\ &= \Delta \langle [\phi_i'] \rangle. \end{aligned} \tag{8.3}$$

As a consequence, the first-order statistics κ vanishes in response to any input pattern orthogonal to $I^{(0)}$.

When activity is readout by the specific decoding vector w, the readout signal takes value:

$$\begin{aligned} z &= \langle w_i[\phi_i]\rangle \\ &= \langle w_i \int \mathcal{D}z\phi(m_i\kappa + I_i + \sqrt{\Delta_0^I}z)\rangle \\ &= \langle w_i \int \mathcal{D}z\phi(w_i\kappa + I_i + \sqrt{\Delta_0^I}z)\rangle \\ &= \kappa\Sigma_w^2\langle[\phi_i']\rangle, \end{aligned} \tag{8.4}$$

while we trivially obtain $z = 0$ for any decoding set orthogonal to both structure vectors m and n.

In conclusion, a non-vanishing readout response requires both an external input correlated with the Go pattern $I^{(0)}$ and a decoding set correlated with the specifically designed readout w.

In Fig. 8.2 **d**, we test the generalization properties of a a network which responds to two Go patterns $I^{(0)}$ and $I^{(1)}$. We examine the response to a normalized mixture input defined as:

$$I = \sqrt{\alpha}I^{(0)} + \sqrt{1-\alpha}I^{(1)}, \tag{8.5}$$

so that the variance of the total input is fixed and equal to Σ_I^2.

Detectors of correlations In Fig. 8.2 **c**, we show that it is possible to obtain highly non-linear readout responses by considering non-vanishing overlaps between the structure vectors m and n. The simplest setup consists of taking:

$$\begin{aligned} m &= w + \rho_m z \\ n &= I^{(0)} + \rho_n z, \end{aligned} \tag{8.6}$$

where z is a standard gaussian vector which defines an additional direction orthogonal both to w and $I^{(0)}$. In this configuration, the structure strength is given by $\rho_m\rho_n$.

As it has been shown in Fig. 7.1, large values of the structure overlap $\rho_m\rho_n$ generate two bistable solutions. If the external input correlates with the preferred one ($\Delta > 0$), two asymmetric solutions exist only when the value of the input Δ is not too large (see Fig. 7.3 **f-g**). In this regime, the two stable values of κ average very close to zero, so that (because of Eq. 8.4) the average readout vanishes. When the correlation Δ is large, instead, only the positive branch of the solution is retrieved (Fig. 7.3 **f-g**). The average value of κ, and thus the readout signal, are positive for every initial condition of the dynamics.

The threshold value for Δ at which the readout value becomes positive is mostly determined by the strength of the structure overlap (see Fig. 7.3 **f**), and depends on the input and readout parameters Σ_I and Σ_w. As it has been shown in Fig. 7.3 **i-j**, moreover, the position of the transition can be further controlled by introducing additional external inputs which correlate with the left-structure vector n on directions that are perpendicular to $I^{(0)}$. Here we adopt modulatory inputs γ that are aligned on the shared z direction:

$$I = I^{(0)}\frac{\Delta}{\Sigma_I^2} + x\sqrt{\Sigma_I^2 - \frac{\Delta^2}{\Sigma_I^4}} + \gamma z. \tag{8.7}$$

For practical purposes, in order to obtain the results of Fig. 8.2 **c**, we first fix the values of Σ_I, Σ_w and ρ_n. We then tune the value of ρ_m in order to obtain a threshold value for Δ

close to 0.5. We finally considered two modulatory inputs of different sign ($\gamma_1 = 0.2$ and $\gamma_2 = -0.2$), that have the effect of moving the value of the threshold in the two different directions.

8.2 Computing with rank two structures

This far we focused on unit rank structure in the connectivity. A more general structured component of rank $r \ll N$ can be written as

$$P_{ij} = \frac{m_i^{(1)} n_j^{(1)}}{N} + \ldots + \frac{m_i^{(r)} n_j^{(r)}}{N}, \tag{8.8}$$

and is in principle characterized by $2r$ vectors $m^{(k)}$ and $n^{(k)}$. Based on the analysis of the unit rank case, we expect that the dynamics of a network with rank r structure to lie in the $r-$dimensional subspace spanned by the r right-structure vectors $m^k, k = 1 \ldots r$. The details of the dynamics will depend on the geometrical arrangement of these $2r$ vectors among themselves and with respect to the input pattern. The number of possible configurations increases very quickly with the structure rank. In the remaining of this manuscript, we will explore only the rank two case, and show that even for $r = 2$ the dynamical and computational repertoire is already rich.

A rank two connectivity structure is fully specified by two right vectors $m^{(1)}$ and $m^{(2)}$, and two left vectors $n^{(1)}$ and $n^{(2)}$:

$$P_{ij} = \frac{m_i^{(1)} n_j^{(1)}}{N} + \frac{m_i^{(2)} n_j^{(2)}}{N}, \tag{8.9}$$

where the vector pairs $m^{(1)}$ and $m^{(2)}$, $n^{(1)}$ and $n^{(2)}$ are assumed to be linearly independent. The activity of the network in response to an input pattern I_i is in general given by

$$\mu_i = \kappa_1 m_i^{(1)} + \kappa_2 m_i^{(2)} + I_i, \tag{8.10}$$

where κ_1 and κ_2 are the projections of average network activity $[\phi]$ on the left-structure vectors $n^{(1)}$ and $n^{(2)}$. The activity evoked in response to an input is therefore in general three-dimensional, and lies in a subspace spanned by the right-structure vectors $m^{(1)}$ and $m^{(2)}$ and the input vector I.

As in the case of unit rank structures, we determine the network statistics by exploiting the link between linear stability analysis and mean field description. The study of the properties of eigenvalues and eigenvectors for the low-dimensional matrix P_{ij} helps to predict the complex behaviour of activity above the instability and to restrict our attention to the cases of interest.

Note that the non-linear network dynamics is determined by the relative orientation of the structure and input vectors, but also by the characteristics of the statistical distribution of their elements. In contrast to the cases we analyzed so far, the distribution of the entries in the structure vectors can play indeed a major role when the rank of P_{ij} is larger than unity. In the following, we focus on the case of broadly, normally distributed patterns.

Throughout the rest of this chapter, we consider several specific rank two configurations which are of interest for computational applications.

8.3 Implementing the 2AFC task

As direct extension of the network architecture presented in Section 8.1, we design a rank two structure which allows multiple specific and selective input-output associations.

We consider the simplest possible rank two connectivity matrix, where the four structure vectors are independently chosen and therefore mutually orthogonal.

As the overlap between left- and right-structure vectors vanishes, in absence of inputs there is no structure in the spontaneous activity. As we formally show in the following, structure in the activity can only be evoked by input patterns that overlap with a left-structure vector. As the left-structure vectors $n^{(1)}$ and $n^{(2)}$ are orthogonal, they select independent inputs and project them onto independent output directions $m^{(1)}$ and $m^{(2)}$. The two unit-rank terms in the connectivity therefore implement two independent input-output channels.

Such a setup for instance allows us to directly extend the implementation of the Go-Nogo task to a two alternative-choice task (2AFC) (Fig. 8.3 **a-b**), in which two different classes of inputs (implemented by $n^{(1)}$ and $n^{(2)}$) are mapped to two different readout directions ($m^{(1)}$ and $m^{(2)}$). Another possibility is that the two input directions represent two different features of the stimulus (e.g. color and motion [83]) that are bound together in the network activity, but can be independently extracted by reading-out along the directions of $m^{(1)}$ and $m^{(2)}$.

8.3.1 Mean field equations

We start by analyzing in detail the input-free dynamics of a network with a rank two structure that has been built from orthogonal structure vectors. Similarly to the unit rank case, if the structure vectors are orthogonal, the network is silent in absence of external inputs: $\kappa^1 = \kappa^2 = 0$. A single homogeneous state – stationary or chaotic – is the unique stable attractor of the dynamics. Consistently, the eigenspectrum of J_{ij} does not contain any outlier, since every eigenvalue of P_{ij} vanishes.

In order to compute the eigenspectrum of P_{ij}, we can rotate the matrix onto a basis defined by an orthonormal set of vectors, and compute its eigenvalues in the transformed basis. For simplicity, we consider an orthonormal set whose first four vectors are built from the structure vectors:

$$\begin{aligned} u_1 &= \alpha_1 m^{(1)} \\ u_2 &= \alpha_2 m^{(2)} \\ u_3 &= \alpha_3 n^{(1)} \\ u_4 &= \alpha_4 n^{(2)}, \end{aligned} \tag{8.11}$$

where the coefficient α_k ($k = 1, ..., 4$) denote the normalization factors. In this basis, the first four rows and columns of the rotated matrix P'_{ij} read:

$$P'_{ij} = \frac{1}{N} \begin{pmatrix} 0 & 0 & \frac{1}{\alpha_1\alpha_3} & 0 \\ 0 & 0 & 0 & \frac{1}{\alpha_2\alpha_4} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{8.12}$$

all the remaining entries being fixed to 0. From the present matrix form, it easy to verify that all the eigenvalues of P'_{ij}, and thus all the eigenvalues of P_{ij}, vanish. Note that rewriting P_{ij} in

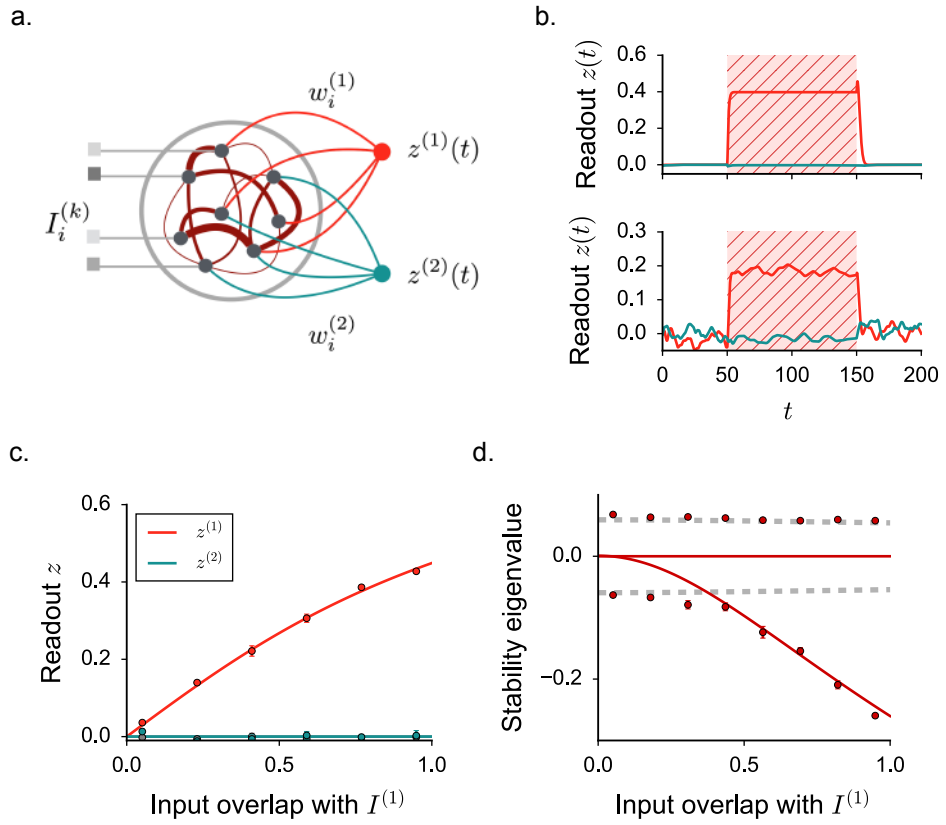


FIGURE 8.3: Rank two structures with null overlap can be used to implement multiple input-output selective associations as in the 2AFC task. **a.** The partially structured reservoir receives an external input pattern $I_i^{(k)}$. The recurrent structure is defined as in Eq. 8.18. **b.** Samples of the readout activity in the stationary and in the chaotic regime. The response of the $z^{(1)}$ (resp. $z^{(2)}$) readout is displayed in red (resp. green). During the shaded time window, the stimulus $I^{(1)}$ is presented to the network, and the readout signal $z^{(1)}$ displays a response. **c.** Readouts values as a function of the value of the overlap between the presented and the preferred input $I^{(1)}$. The input overlap values are normalized by the total input variance Σ_I. Grey: activity is decoded from an additional random and orthogonal decoding set. Continuous lines: DMF prediction, dots: average response over $N_{tr} = 6$ networks of size $N = 2000$. **d.** Outliers in the stability eigenspectrum as a function of the overlap Σ_{nI}. Note that one outlier vanishes for every value of the input overlap. The grey dashed line indicate the value of the radius of the compact part of the eigenspectrum. Red dots: real part of the smallest and largest eigenvalues in the spectrum obtained from numerical simulations. Outliers can only be measured numerically when their value is larger than the radius in absolute value. In this panel: $g = 0.1$. Choice of the parameters: $g = 0.8$, $\Sigma_I = 1.4$, $\Sigma_w = 1.2$.

an orthonormal basis simplifies the search for its eigenvalues also in more complex cases where the structure vectors share several overlap directions. In those cases, a proper basis needs to be built starting from the structure vectors through a Gram-Schmidt orthonormalization process.

As a side note we observe that, even though P'_{ij} (and thus P_{ij}) admits only vanishing eigenvalues, its rank is still equal to two. Indeed, the rank can be computed as N minus the dimensionality of the kernel associated to P'_{ij}, defined by any vector x obeying $P'x = 0$. As P'_{ij} contains $N-2$ empty rows, the last equations imposes two independent contraints on the components of x. As a consequence, the dimensionality of the kernel equals $N-2$, and the rank is equal to two.

We turn to consider the non-trivial responses that are obtained in presence of external inputs. We examine the network dynamics in response to an input $\tilde{I}$ which partially correlates with one of the left-structure vectors, here $n^{(1)}$ (see Eq. 8.2):

$$\tilde{I} = n^{(1)} \frac{\Sigma_{nI}}{\Sigma_I^2} + x \sqrt{\Sigma_I^2 - \frac{\Sigma_{nI}^2}{\Sigma_I^4}}. \tag{8.13}$$

Similarly to the unit rank case, we find that $\tilde{I}$ elicits a network response in the plane $\tilde{I} - m^{(1)}$. The overlap values are indeed given by:

$$\begin{aligned} \kappa_1 &= \Sigma_{nI} \langle [\phi'_i] \rangle \\ \kappa_2 &= 0, \end{aligned} \tag{8.14}$$

and they can be used to close the mean field equations together with the equation for the first ($\mu = 0$) and second-order statistics. In the case of stationary states we have:

$$\Delta_0 = g^2 \langle [\phi_i^2] \rangle + \Sigma_m^2 \left(\kappa_1^2 + \kappa_2^2 \right) + \Sigma_I^2. \tag{8.15}$$

Similar arguments allow to derive the two equations needed for the chaotic states.

In order to assess the stability of the stationary states, we extend the procedure illustrated in Chapter 7 and we evaluate the position of the outliers in the stability eigenspectrum by computing the eigenvalues of a reduced stability matrix $\mathcal{M}$. The step-by-step derivation of $\mathcal{M}$ in the case of generic rank two structures is given in Appendix F. The result can be written as:

$$\mathcal{M} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2g^2 \langle [\phi_i \phi'_i] \rangle & g^2 \left\{ \langle [\phi_i'^2] \rangle + \langle [\phi_i \phi''_i] \rangle \right\} & 2\Sigma_m^2 \kappa_1^0 & 2\Sigma_m^2 \kappa_2^0 \\ 2b_1 g^2 \langle [\phi_i \phi'_i] \rangle & b_1 g^2 \left\{ \langle [\phi_i'^2] \rangle + \langle [\phi_i \phi''_i] \rangle \right\} & 2b_1 \Sigma_m^2 \kappa_1^0 + a_{11} & 2b_1 \Sigma_m^2 \kappa_2^0 + a_{12} \\ 2b_2 g^2 \langle [\phi_i \phi'_i] \rangle & b_2 g^2 \left\{ \langle [\phi_i'^2] \rangle + \langle [\phi_i \phi''_i] \rangle \right\} & 2b_2 \Sigma_m^2 \kappa_1^0 + a_{21} & 2b_2 \Sigma_m^2 \kappa_2^0 + a_{22} \end{pmatrix}. \tag{8.16}$$

In the case of orthogonal structures and correlated input patterns $\tilde{I}$, a little algebra reveals that all the a values vanish, while we have:

$$\begin{aligned} b^1 &= \frac{1}{2} \Sigma_{nI} \langle [\phi''_i] \rangle \\ b^2 &= 0. \end{aligned} \tag{8.17}$$

We conclude that the first and the last row of $\mathcal{M}$ always vanish. Furthermore, the second and the third rows are proportional one to the other. As a consequence, the stability analysis predicts at most one outlier eigenvalue, which is indeed observed in the spectrum (Fig. 8.3 **d**).

The outlier is negative, as the effect of introducing inputs in the direction of the left vector $n^{(1)}$ is to further stabilize the dynamics. As it will be shown, more than one outlier can be observed in the case where the low-dimensional structure involves overlap directions.

To conclude, we discuss how orthogonal rank two structures can be used to build up a network implementation for the two-alternative forced choice (2AFC) task. Let us consider again a model network which receives one among several orthogonal input patterns $I^{(k)}$. The network is provided with two output readout signals, defined as: $z^{(1)} = \langle w^{(1)}[\phi_i]\rangle$ and $z^{(2)} = \langle w^{(2)}[\phi_i]\rangle$, where $w^{(1)}$ and $w^{(2)}$ are two orthogonal readout sets. We want the network to associate a response in $z^{(1)}$ (resp. $z^{(2)}$) every time an input pattern which is partially correlated to $I^{(1)}$ (resp. $I^{(2)}$) is presented. Similarly to Fig. 8.1, the simplest structure which correctly implements the task is given by:

$$\begin{aligned} m^{(1)} &= w^{(1)} \\ m^{(2)} &= w^{(2)} \\ n^{(1)} &= I^{(1)} \\ n^{(2)} &= I^{(2)}. \end{aligned} \tag{8.18}$$

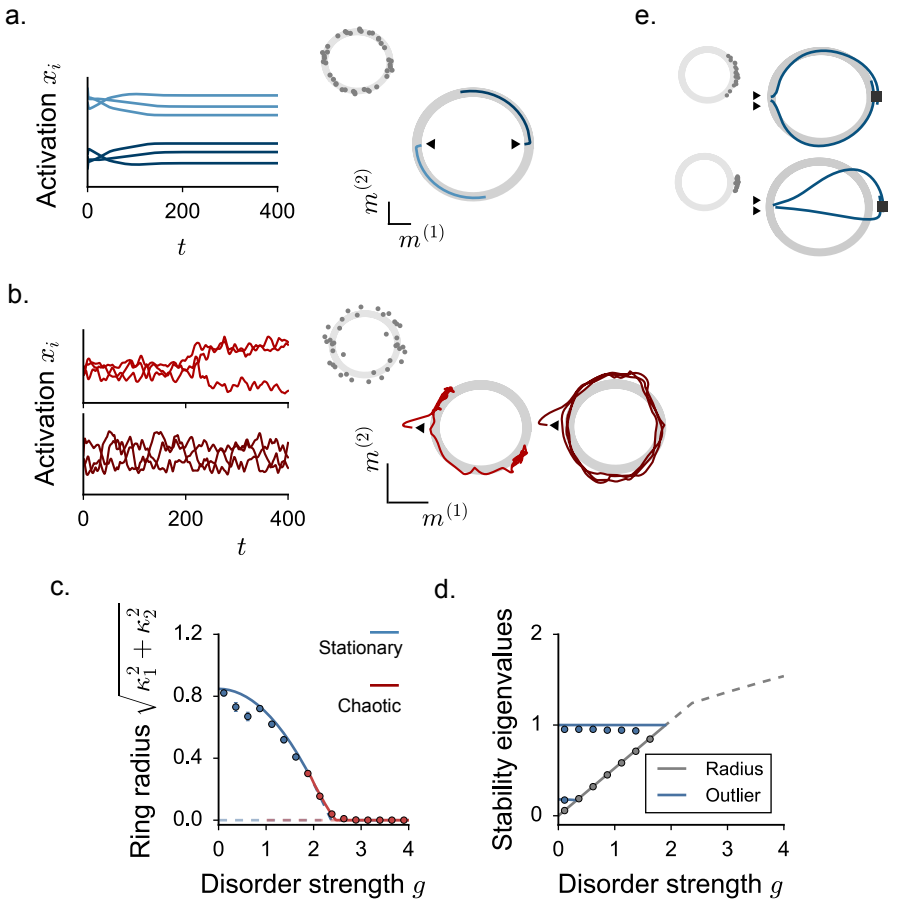
The resulting selectivity and specificity properties of the network response are illustrated in Fig. 8.3 **b-c**.

8.4 Building a ring attractor

Overlaps between different structure vectors give rise to more complex spontaneous dynamics, and a richer range of responses to external inputs. As a direct extension of the unit rank case, we consider next the situation where the pairs $m^{(1)} - n^{(1)}$ and $m^{(2)} - n^{(2)}$ each share a different common direction, the corresponding overlaps being $\rho_1 = m^{(1)T} n^{(1)}/N$ and $\rho_2 = m^{(2)T} n^{(2)}/N$. As in the unit rank case, each of these overlaps can lead to bistable, structured spontaneous activity, so that in general there will be four structured spontaneous states (two states with $\kappa_1 \neq 0, \kappa_2 = 0$ for $\rho_1 > 1$ and two states with $\kappa_1 = 0, \kappa_2 \neq 0$ for $\rho_2 > 1$).

A particularly interesting situation however occurs when the two unit rank contributions are symmetric, so that the two overlaps are equal, $\rho_1 = \rho_2 = \rho$. In that case, our theoretical analysis predicts the existence of a continuum of structured spontaneous states that explore a two-dimensional circle in the $m^{(1)} - m^2$ plane (Fig. 8.4), i.e. a so-called ring attractor [17], embedded in the N-dimensional space of activity. That theoretical prediction formally holds in the limit of infinite-size networks; in simulations of finite size networks, the dynamics instead always converge on a small number of equilibrium spontaneous states located on the ring [10, 29]. The equilibrium reached in a given situation is determined by the corresponding realization of the random part of the connectivity, and the initial conditions. Different realizations of the random connectivity lead to different equilibrium states, which all however lie on the predicted ring (Fig. 8.4 **a**). For a given realization of the random connectivity, transient dynamics moreover show a clear signature of the ring structure. Indeed the points on the ring are close to stable and form a slow manifold (Fig. 8.4 **a**). The convergence to the equilibrium activity is therefore very slow, and the the temporal dynamics explore the ring structure.

The ring structure in the dynamics is remarkably robust with respect to the random part of the connectivity, which affects it in two different manners (Fig. 8.4 **c-d**). First, increasing the disorder strength g eventually leads to chaotic activity as in the unit-rank case. Second,



increasing g decreases the radius of the ring until it vanishes. Interestingly, the onset of chaos takes place before the ring vanishes, so that when structure and disorder have comparable strengths, chaotic states with ring structure appear. Simulations of finite size networks show that in this situation, the chaotic dynamics are low-dimensional and either explore the whole ring structure, or jump between two states along the ring (Fig. 8.4 **b**).

An external input along a given direction in $n^{(1)} - n^{(2)}$ plane will in general eliminate the continuum of ring solutions, and stabilize the dynamics along one the corresponding direction in the $m^{(1)} - m^{(2)}$ plane (Fig. 8.4 **e**). If the input is weak, although our theory predicts a single stable solution, finite-size simulations still show clear signatures of the underlying ring attractor, as the equilibrium states still depend on the realization of the random connectivity, and the transients display slow dynamics along the ring (Fig. 8.4 **e**, top). If the input is strong, only one equilibrium state persists, the transients are faster and do not necessary lie along the ring (Fig. 8.4 **e**, bottom). Finally inputs that are orthogonal to the $n^{(1)} - n^{(2)}$ plane preserve the ring structure, and only modulate its radius.

Overlaps within the $m^{(1)} - n^{(1)}$ and $m^{(2)} - n^{(2)}$ pairs in a rank two structure therefore lead to novel dynamical phenomena with respect to a unit rank structure, and in particular ring attractors that have been implicated in modelling a range of experimental phenomena

Figure 8.4 *(previous page)*: Ring attractor from rank two connectivity structure. **a.** Sample of activity from a finite-size realization ($N = 4000$) of the structured connectivity matrix. Activity is initialized in two different initial conditions (light and dark blue), indicated by the small arrows. Left: time traces of the activation variables for three randomly selected network units. Note the long time range on the x axis. Right: population activity projected on the plane spanned by the right vectors $m^{(1)}$ and $m^{(2)}$. The ring solution predicted by the mean field theory is displayed in light gray. The strength of the disorder is $g = 0.5$, so that the network is in a stationary regime. In the small inset, we reproduce the theoretical prediction together with the final state of additional $N_{tr} = 20$ networks realizations, that are displayed as grey dots. **b.** Sample of activity for two finite-size realizations ($N = 4000$) of the structured connectivity matrix (light and dark red). Details as in **a**. The strength of random connections is $g = 2.1$, so that the network is in a chaotic regime. Chaotic fluctuations can occur together with a slow exploration of the ring (light red). If two specific states on the ring appear to be more stable, chaotic fluctuations can induce jumps between the two of them (dark red). **c-d.** Mean field characterization of the ring structure: radius of the ring attractor and stability eigenvalues. All the details are as in Fig. 7.2. Dots: numerical results from finite-size ($N = 5000$) networks, average over 10 realizations of the connectivity matrix. **e.** Input response for two finite-size networks. Input patterns which correlate with the left vector $n^{(1)}$ reduce the ring attractor to a single stable state (black square). Activity is thus projected in the direction spanned by the right vector $m^{(1)}$. The grey ring displays the mean field solution in absence of external inputs ($g = 0.5$, as in **a**). In the top panel, the input is weak ($\Sigma_I = 0.2$). The transient dynamics as well as the equilibrium state lie close to the ring structure. In the bottom panel, the input is strong ($\Sigma_I = 0.6$), and the ring structure is not anymore apparent. Figure details as in **a-b**, with $\Sigma = 2.0$, $\rho_1 = \rho_2 = 1.6$.

such as orientation selectivity [17], grid cells [29] and working memory [39]. More generally each of the pairs out of the four vectors $m^{(1)}, n^{(1)}, m^{(2)}, n^{(2)}$ may share a common direction. Instead of attempting to map all existing possibilities, in the next two sections we describe two particularly interesting setups.

8.4.1 Mean field equations

We provide here a quantitative characterization of the partially structured networks that have been used to build continuous ring attractors. We consider rank two structures where the two structure pairs $m^{(1)}$ and $n^{(1)}$, $m^{(2)}$ and $n^{(2)}$ share two different overlap directions, defined by vectors y_1 and y_2. We set:

$$\begin{aligned} m^{(1)} &= \sqrt{\Sigma^2 - \rho_1^2}\, x_1 + \rho_1 y_1 \\ m^{(2)} &= \sqrt{\Sigma^2 - \rho_2^2}\, x_2 + \rho_2 y_2 \\ n^{(1)} &= \sqrt{\Sigma^2 - \rho_1^2}\, x_3 + \rho_1 y_1 \\ n^{(2)} &= \sqrt{\Sigma^2 - \rho_2^2}\, x_4 + \rho_2 y_2. \end{aligned} \tag{8.19}$$

where Σ^2 is the variance of the structure vectors and ρ_1^2 and ρ_2^2 quantify the overlaps along the directions y_1 and y_2.

By rotating P_{ij} onto the orthonormal basis that can be built from $m^{(1)}$ and $m^{(2)}$ by orthogonalizing the left vectors $n^{(1)}$ and $n^{(2)}$, one can easily check that the two non-zero eigenvalues of P_{ij} are given by $\lambda_1 = \rho_1^2$ and $\lambda_2 = \rho_2^2$. They correspond, respectively, to the two right-eigenvectors $m^{(1)}$ and $m^{(2)}$. In absence of external inputs, an instability is thus likely to occur in the direction of the $m^{(k)}$ vector which corresponds to the strongest overlap.

The specific case we consider in Fig. 8.4 corresponds to the degenerate condition where the two overlaps are equally strong, $\rho_1 = \rho_2 = \rho$, and any combination of $m^{(1)}$ and $m^{(2)}$ is a right-eigenvector. The mean field equations for the first-order statistics read:

$$\begin{aligned} \kappa_1 &= \rho^2 \kappa_1 \langle [\phi_i'] \rangle \\ \kappa_2 &= \rho^2 \kappa_2 \langle [\phi_i'] \rangle. \end{aligned} \tag{8.20}$$

Similarly to Eq. 7.78, the two equations admit a silent ($\kappa_1 = \kappa_2 = 0$) and a non-trivial state, determined by two identical conditions which read:

$$1 = \rho^2 \langle [\phi_i'(0, \tilde{\Delta}_0)] \rangle. \tag{8.21}$$

The equation above determines the value of $\Delta_0 = \tilde{\Delta}_0$. Note that the non-trivial state exists only for $\rho > 1$.

A second condition is imposed by the equation for the second-order momentum which reads, for stationary solutions:

$$\Delta_0 = g^2 \langle [\phi_i^2] \rangle + \Sigma^2 \left(\kappa_1^2 + \kappa_2^2 \right). \tag{8.22}$$

As the value of Δ_0 is fixed, the mean field set of equations fixes only the sum $\kappa_1^2 + \kappa_2^2$, but not each single component. The mean field thus returns a one-dimensional continuum of solutions, the shape of which resembles a ring of radius $\sqrt{\kappa_1^2 + \kappa_2^2}$ in the $m^{(1)} - m^{(2)}$ plane (see Fig. 8.4 **a-b**). Similarly to the unit rank case, the value of the radius can be computed explicitly by solving numerically the two mean field equations (three in the case of chaotic regimes), and depends on the relative magnitude of ρ^2 compared to g (Fig. Fig. 8.4 **c**). Highly disordered connectivities have the usual effect of suppressing non-trivial structured solutions in favour of homogeneous and unstructured states. For sufficiently high g values, furthermore, structured solution can display chaotic dynamics (Fig. 8.4 **c**, red).

A linear stability analysis reveals that the one-dimensional solution consists of a continuous set of marginally stable states. Similarly to the orthogonal vectors case, the position of the outliers in the eigenspectra of S_{ij} can be evaluated by computing the reduced stability matrix $\mathcal{M}$, which reads:

$$\mathcal{M} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2g^2 \langle [\phi_i \phi_i'] \rangle & g^2 \left\{ \langle [\phi_i'^2] \rangle + \langle [\phi_i \phi_i''] \rangle \right\} & 2\Sigma_m^2 \kappa_1^0 & 2\Sigma_m^2 \kappa_2^0 \\ 2b_1 g^2 \langle [\phi_i \phi_i'] \rangle & b_1 g^2 \left\{ \langle [\phi_i'^2] \rangle + \langle [\phi_i \phi_i''] \rangle \right\} & 2b_1 \Sigma_m^2 \kappa_1^0 + a_{11} & 2b_1 \Sigma_m^2 \kappa_2^0 \\ 2b_2 g^2 \langle [\phi_i \phi_i'] \rangle & b_2 g^2 \left\{ \langle [\phi_i'^2] \rangle + \langle [\phi_i \phi_i''] \rangle \right\} & 2b_2 \Sigma_m^2 \kappa_1^0 & 2b_2 \Sigma_m^2 \kappa_2^0 + a_{22} \end{pmatrix}, \tag{8.23}$$

with:

$$\begin{aligned} a_{11} &= \rho^2 \langle [\phi_i'] \rangle \\ b_1 &= \frac{1}{2} \rho^2 \kappa_1^0 \langle [\phi_i'''] \rangle \end{aligned} \tag{8.24}$$

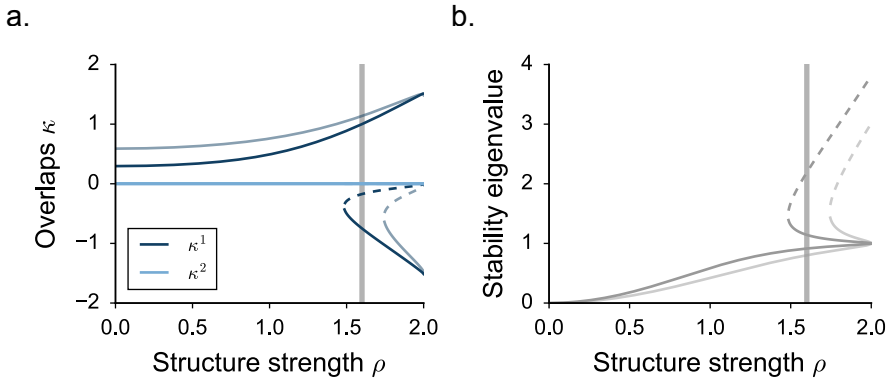


FIGURE 8.5: Mean field solutions for network models with rank two structures characterized by pairwise, internal overlaps ρ^2. Response to the input pattern $\tilde{I}$, which correlates with the component x_3 of the left-structure vector $n^{(1)}$ ($\alpha = 0$ in Eq. 8.26). **a.** Values of the overlaps κ_1 and κ_2 as a function of the parameter ρ, which controls the structure strength. Stable solutions are plotted as continuous lines, unstable ones as dashed. Solid (resp. transparent) lines refer to weak (resp. strong) external inputs: $\Sigma_I = 0.2$ (resp. 0.6). The vertical gray line indicate the value of ρ that has been used in Fig. 8.4. **b.** Value of the largest outlier in the eigenspectrum of the linear stability matrix S_{ij}, computed from the reduced matrix $\mathcal{M}$ (Eq. 8.23). Note that only one branch of the solution (the one corresponding to positive κ_1 values) is stable. Choice of the parameters: $\Sigma = 2.0$, $\rho = 1.6$, $g = 0.5$.

and

$$\begin{aligned} a_{22} &= \rho^2 \langle [\phi_i'] \rangle \\ b_2 &= \frac{1}{2}\rho^2 \kappa_2^0 \langle [\phi_i'''] \rangle. \end{aligned} \tag{8.25}$$

As shown in Fig. 8.4 **d**, diagonalizing the stability matrix $\mathcal{M}$ returns the values of two distinct outlier eigenvalues. The third non-zero eigenvalue of $\mathcal{M}$ lays instead systematically inside the compact component of the spectrum, and corresponds to an average measure of the time scales inherited by the random modes. One of the two outliers is tuned exactly on the stability boundary for every value of the parameters which generate a ring solution. This marginally stable eigenvalue is responsible for the slow dynamical time scales which are observed in numerical simulations of the network activity (Fig. 8.4 **a-b**).

We next examine how the structured, ring-shaped solution is perturbed by the injection of external input patterns.

We consider an input pattern $\tilde{I}$ of variance Σ_I^2. When $\tilde{I}$ does not share any overlap direction with the left vectors $n^{(1)}$ and $n^{(2)}$, the mean field equations are affected solely by an extra term Σ_I which needs to be included in the equation for the second-order statistics (Eq. 8.22). As the equations for the first-order statistics do not change, the one-dimensional degeneracy of the solution persists. The extra term Σ_I^2 however decreases the value of the radius of the ring.

When the input contains a component which overlaps with one or both left vectors $n^{(1)}$ and $n^{(2)}$, the degeneracy in the two equations for κ_1 and κ_2 is broken. As a consequence, the one-dimensional solution collapses onto a unique stable point. Consider for example an input

pattern of the form:

$$\tilde{I} = \Sigma_I \left(\sqrt{1-\alpha}\, x_3 + \sqrt{\alpha}\, x_4 \right). \tag{8.26}$$

The equations for the first order become:

$$\begin{aligned} \kappa_1 &= \left(\rho^2 \kappa_1 + \Sigma_I \sqrt{1-\alpha} \sqrt{\Sigma^2 - \rho^2} \right) \langle [\phi_i'] \rangle \\ \kappa_2 &= \left(\rho^2 \kappa_2 + \Sigma_I \sqrt{\alpha} \sqrt{\Sigma^2 - \rho^2} \right) \langle [\phi_i'] \rangle \end{aligned} \tag{8.27}$$

or, alternatively:

$$\begin{aligned} \kappa_1 &= \frac{\Sigma_I \sqrt{1-\alpha} \sqrt{\Sigma^2 - \rho^2} \langle [\phi_i'] \rangle}{1 - \rho^2 \langle [\phi_i'] \rangle} \\ \kappa_2 &= \frac{\Sigma_I \sqrt{\alpha} \sqrt{\Sigma^2 - \rho^2} \langle [\phi_i'] \rangle}{1 - \rho^2 \langle [\phi_i'] \rangle}. \end{aligned} \tag{8.28}$$

The values of κ_1 and κ_2 are thus uniquely specified, and can be computed by iterating the two equations together with the expression for the second-order statistics:

$$\Delta_0 = g^2 \langle [\phi_i^2] \rangle + \Sigma^2 \left(\kappa_1^2 + \kappa_2^2 \right) + \Sigma_I^2. \tag{8.29}$$

In a similar way, the presence of correlated external inputs affect the values of the entries of the reduced stability matrix $\mathcal{M}$:

$$\begin{aligned} b_1 &= \frac{1}{2} \left(\rho^2 \kappa_1^0 + \Sigma_I \sqrt{1-\alpha} \sqrt{\Sigma^2 - \rho^2} \right) \langle [\phi_i'''] \rangle \\ b_2 &= \frac{1}{2} \left(\rho^2 \kappa_2^0 + \Sigma_I \sqrt{\alpha} \sqrt{\Sigma^2 - \rho^2} \right) \langle [\phi_i'''] \rangle. \end{aligned} \tag{8.30}$$

In Fig. 8.4 and 8.5, we focus on the case of an external input pattern aligned with x_3 (and thus $n^{(1)}$) . We fix $\alpha = 0$, that implies $\kappa_2 = 0$.

Solving the mean field equations reveal that, according to the strength of the input Σ_I, one or three fixed points exist. When the input is weak with respect to the structure overlap ρ^2, two fixed points appear in the proximity of the ring, along the direction defined by the axis $\kappa^2 = 0$ (Fig. 8.4 **e** top and Fig. 8.5 **a**). In particular, when $\tilde{I}$ positively correlates with $n^{(1)}$, only the fixed point with positive value of κ_1 gets stabilized. The remaining two solutions are characterized by one outlier eigenvalue which lays above the instability boundary, and are thus unstable (Fig. 8.5 **a-b**). On the other hand, when the input is sufficiently strong, solely the stable fixed point survives (Fig. 8.4 **e** bottom and Fig. 8.5 **a-b**). Activity is then robustly projected in the direction defined by the right vector $m^{(1)}$.

8.5 Implementing a context-dependent discrimination task

To illustrate the computational capacity of networks with rank two structures, here we exploit them to implement a complex behavioral task, a context-dependent decision making paradigm inspired by a non-human primate study [83].

We consider a situation where the stimuli consist of combinations of two different features A and B. In the experimental study [83], the stimuli were random dot kinetograms, and the features A and B correspond to the direction of motion and color of these stimuli. The tasks

consists in classifying the stimuli according to one of those features, the relevant one being indicated by an explicit contextual cue.

Population recordings in the prefrontal cortex showed that the relevant stimulus feature was not pre-selected in the sensory areas, but that both motion and color signals were present in the prefrontal cortex independently of the contextual cue. The representation of these signals was highly mixed and distributed over the recorded PFC population, but could be captured by two main directions in the neural space. Following these experimental observations and the model used in [83], we represented the pattern of inputs to the network on a given trial as a vector I given by (Fig. 8.6):

$$I = c_A I_A + c_B I_B + \gamma_A I_{ctxA} + \gamma_B I_{ctxB} + \text{noise}. \tag{8.31}$$

Here c_A and c_B are two scalar values that represent the strengths of features A and B in the presented stimulus, while γ_A and γ_B are two binary values that represent the presence or absence of the cues for context A and B. I_A, I_B, I_{ctxA} and I_{ctxB} are N-dimensional vectors that represent the directions of population inputs corresponding to stimulus features and contextual cues. These vectors are generated randomly and fixed, while the four scalars c_A, c_B, γ_A and γ_B vary from trial to trial. Note that the two stimulus features are represented as being orthogonal, so that the sensory stimuli cover a portion of the two-dimensional subspace spanned by I_A and I_B. The contextual cues increase the dimensionality of the total input to the network to four.

We implemented a Go-Nogo version of the task, in which the output is required to be non-zero when the relevant feature is stronger than a prescribed threshold (arbitrarily set to 0.5).

The output of the network is determined through a linear readout of the network activity (Fig. 8.6). A key experimental observation is that the direction of the population readout does not depend on the contextual cue. Following [83], we will therefore represent the linear readout as a fixed random vector w. As both the readout direction w and input directions I_A, I_B, I_{ctxA} and I_{ctxB} have been generated randomly, individual neurons represent complex mixtures of stimulus, context and choice signals as observed experimentally.

The crux of this computational task is that on every trial the irrelevant feature of the task needs to be ignored, even if it is stronger than the relevant feature (e.g. color coherence stronger than motion coherence on a motion trial). The central difficulty is that that the readout is context-independent. Without this constraint, two orthogonal readouts could be used to select independently the two orthogonal features, and the task could be implemented as a relatively straight-forward extension of the one-dimensional Go-Nogo detection task (Fig. 8.3). We will nevertheless show that the context-dependent task can be implemented using single readout by exploiting non-linear gating mechanisms to select the relevant feature. This implementation is achieved with a rank two structure constructed using vectors that have a geometrical arrangement with a direct interpretation in terms of input and readout vectors.

As described above, a rank-two connectivity matrix is specified by two right-structure vectors $m^{(1)}$ and $m^{(2)}$ and two left-structure vectors $n^{(1)}$ and $n^{(2)}$. As shown earlier, the right-structure vectors determine the output of network dynamics and can be used to generate the required readout. We will therefore use two vectors $m^{(1)}$ and $m^{(2)}$ that have a common component along the readout vector w. The left-structure vectors $n^{(1)}$ and $n^{(2)}$ select the input patterns that lead to outputs along $m^{(1)}$ and $m^{(2)}$. We want these two vectors to pick up respectively the features A and B of the stimulus, $n^{(1)}$ and $n^{(2)}$ therefore need to have components along respectively the directions of I_A and I_B. Finally, the contextual inputs need

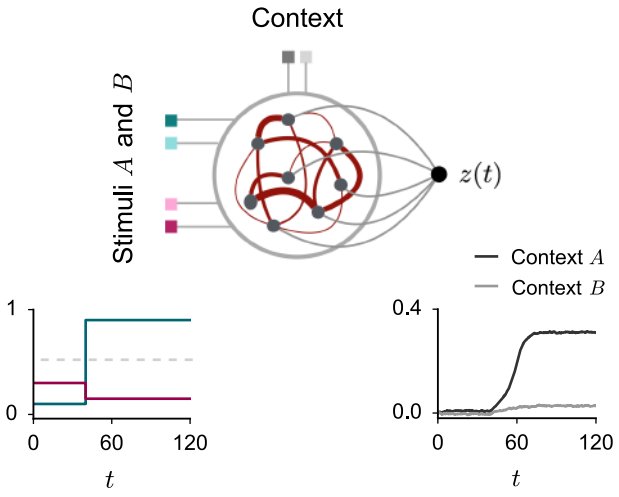


Figure 8.6: Implementing context-dependent computations by using rank two structured connectivities: a two-features discrimination task. The model network receives a mixture of inputs along four main directions (Eq. 8.31): two feature-specified directions I_A and I_B, and two orthogonal directions that are used to modulate the network response in a contex-dependent fashion. The network output z is linearly extracted from the network activity through a single readout set w. The task consists of detecting the presence of strong A (resp. B) inputs during a context A (resp. B) trials, while ignoring the non-relevant feature B (resp. A). In this example (bottom): the readout signal does not respond before $T = 40$, since both input strengths are low. At later times, when the strength of input A becomes large, the network responds only during context A trials, while it remains silent if context B is selected.

to induce a non-linear gating of the selected signals. As shown in Fig. 8.2 **c**, in a unit-rank structure, such gating can be implemented using input along the common direction between left- and right-structure vectors. We therefore add common components to $m^{(1)} - n^{(1)}$ and $m^{(2)} - n^{(2)}$ along respectively the motion and color context vectors I_{ctxA} and I_{ctxB}. The final rank-two setup is described in detail in the next section.

Comparing theoretical predictions and simulations shows that the constructed network performs the required context-dependent computation (Fig. 8.7). Depending on the contextual cue, the output is produced based on only one of the two orthogonal features: in context A, the output is independent of the values of feature B, and conversely in context B. The output therefore behaves as if it were based on two orthogonal readout directions, yet the readout direction is unique and fixed. The context dependent output relies instead on a context-dependent selection of the relevant input features. Such a mechanism was previously suggested based on experimental data by reverse-engineering a trained recurrent network [83]. Here we show that it can be implemented by relying on non-linear dynamics in a network with rank two connectivity structure. Strikingly, the dynamics of the constructed network lie close to a continuum, ring attractor, similarly to what was found in trained recurrent networks [83].

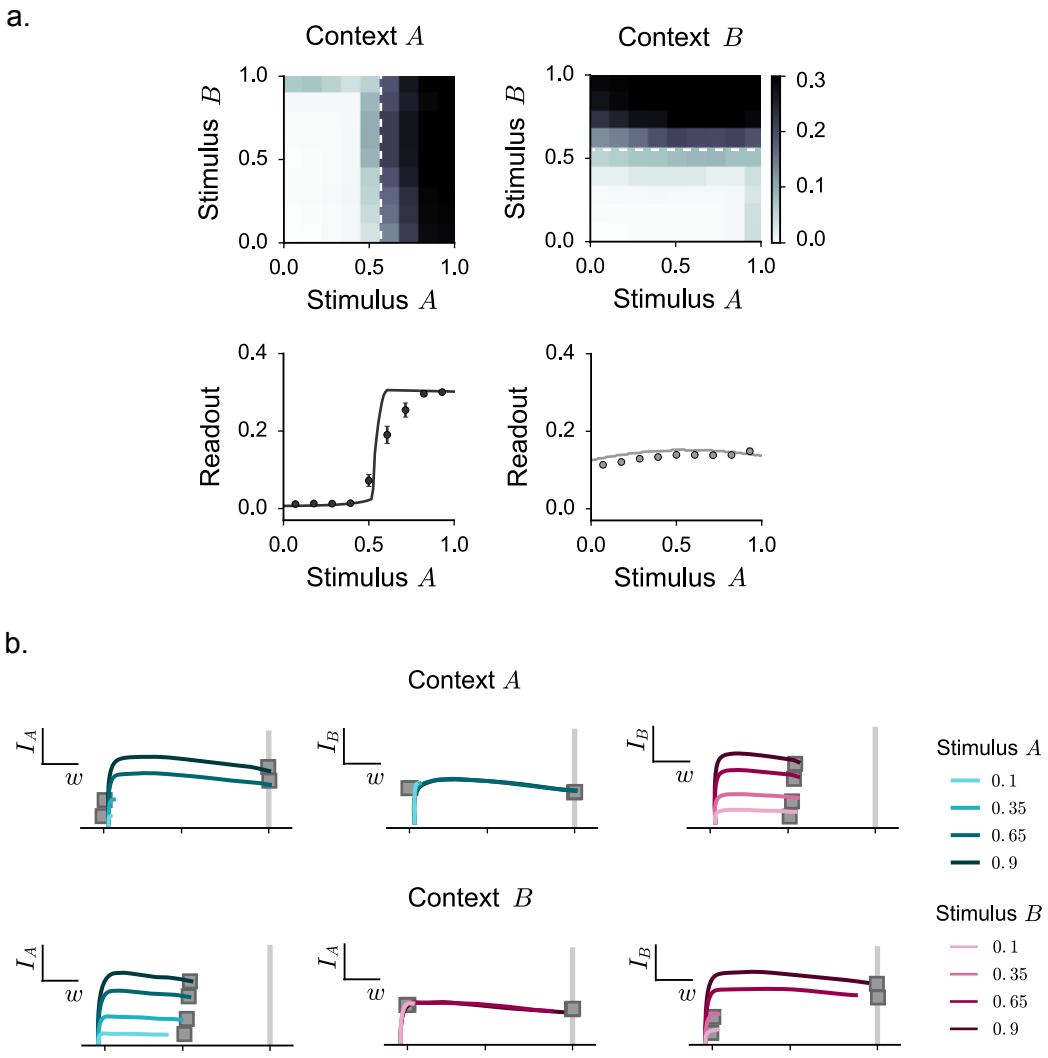


FIGURE 8.7: Implementing context-dependent computations by using rank two structured connectivities: performance. **a.** Top row: readout network response for increasing values of the input strengths along the stimuli directions I_A and I_B. The colormap shows the results from simulated activity in finite-size networks of size $N = 6500$, averaged over 50 different realizations. The mean field theory predicts, on average, high readout values above the the threshold value indicated by the white dashed line. Left and right plots show results for the two different contexts A and B. Bottom row: readout response averaged over different values of feature B. The theoretical prediction is displayed as continuous line, the simulated data as dots. **b.** Time-dependent population activity, projected on the directions that are more salient to the task, as indicated in the small insets. Simulated activity for a single network realization; average over random initial conditions. Top (resp. bottom) row: activity during context A (resp. B) trials. The green (resp. magenta) trajectories refer to network activity for different values of the A (resp. B) feature, averaged across the B (resp. A) feature. The grey dots indicate the theoretical prediction for the steady-state readout values. In this figure: $\Sigma_I = \Sigma_w = 1.2$, $\rho_m = 1.45$, $\rho_n = 3$, $\beta_m = 0.6$, $\beta_n = 1$ (see Section 8.5.1). The values of γ_A and γ_B are fixed to $[0.08, -0.14]$ (resp. $[-0.14, 0.08]$) during the context A (resp. context B) trials.

8.5.1 Mean field equations

Here we provide details on the rank two implementation of the context-dependent discrimination task. The stimuli consist of combination of two different features A and B that correspond to inputs along two directions I_A and I_B. Contextual cues are represented as additional inputs along directions I_{ctxA} and I_{ctxB}. The total input pattern to the network on a given trial is therefore given by

$$\tilde{I} = \frac{\Delta_A}{\Sigma_I^2} I_A + \sqrt{\Sigma_I^2 - \left(\frac{\Delta_A}{\Sigma_I^2}\right)^2} x_1 + \gamma_A I_{ctxA} + \frac{\Delta_B}{\Sigma_I^2} I_B + \sqrt{\Sigma_I^2 - \left(\frac{\Delta_B}{\Sigma_I^2}\right)^2} x_2 + \gamma_B I_{ctxB}. \tag{8.32}$$

The values Δ_A and Δ_B express the strength of the signal along the two input directions. The vectors x_1 and x_2 are two noise terms, while γ_A and γ_B control the two modulatory inputs which are taken in the normalized directions defined by I_{ctxA} and I_{ctxB}. In Fig. 8.7 and 8.8, we indicate with c_A and c_B the normalized strengths Δ_A/Σ_I^2 and Δ_B/Σ_I^2.

In order to design a suitable rank two connectivity matrix, we directly extended the framework that has been used to obtain non-linear outputs in a detection task (Fig. 8.2 **c**). We set:

$$\begin{aligned} m^{(1)} &= y^{(A)} + \rho_m I_{ctxA} \\ n^{(1)} &= I_A + \rho_n I_{ctxA} \\ m^{(2)} &= y^{(B)} + \rho_m I_{ctxB} \\ n^{(2)} &= I_B + \rho_n I_{ctxB}. \end{aligned} \tag{8.33}$$

Note that, because the only overlap directions (I_{ctxA} and I_{ctxB}) are internal to the $m^{(1)} - n^{(1)}$ and $m^{(1)} - n^{(1)}$ pairs, Eq. 8.33 describes a rank two structure which generates a continuous ring attractor as in Fig. 8.4.

In order to implement detection in a context-dependent way, we define a unique readout signal $z(t)$ by using a common readout set w:

$$z = \langle w_i[\phi_i] \rangle. \tag{8.34}$$

The readout $z(t)$ should detect the presence of both stimuli directions. As a consequence, it should be sensitive to both overlap values κ_1 and κ_2. For this reason, we introduce a common term in the four structure vectors that is aligned to the common readout. We obtain:

$$\begin{aligned} m^{(1)} &= y^{(A)} + \rho_m I_{ctxA} + \beta_m w \\ n^{(1)} &= I_A + \rho_n I_{ctxA} + \beta_n w \\ m^{(2)} &= y^{(B)} + \rho_m I_{ctxB} + \beta_m w \\ n^{(2)} &= I_B + \rho_n I_{ctxB} + \beta_n w. \end{aligned} \tag{8.35}$$

Introducing a common overlap direction has the effect of destabilizing the continuous attractor dynamics along the direction $\kappa_1 = \kappa_2$, where two stable and symmetric fixed points are generated. The equations for the first-order input-free dynamics read indeed:

$$\begin{aligned} \kappa_1 &= \langle n^{(1)}[\phi_i] \rangle = \rho_m \rho_n \kappa_1 \langle [\phi_i'] \rangle + \beta_m \beta_n (\kappa_1 + \kappa_2) \langle [\phi_i'] \rangle \\ \kappa_2 &= \langle n^{(2)}[\phi_i] \rangle = \rho_m \rho_n \kappa_2 \langle [\phi_i'] \rangle + \beta_m \beta_n (\kappa_1 + \kappa_2) \langle [\phi_i'] \rangle \end{aligned} \tag{8.36}$$

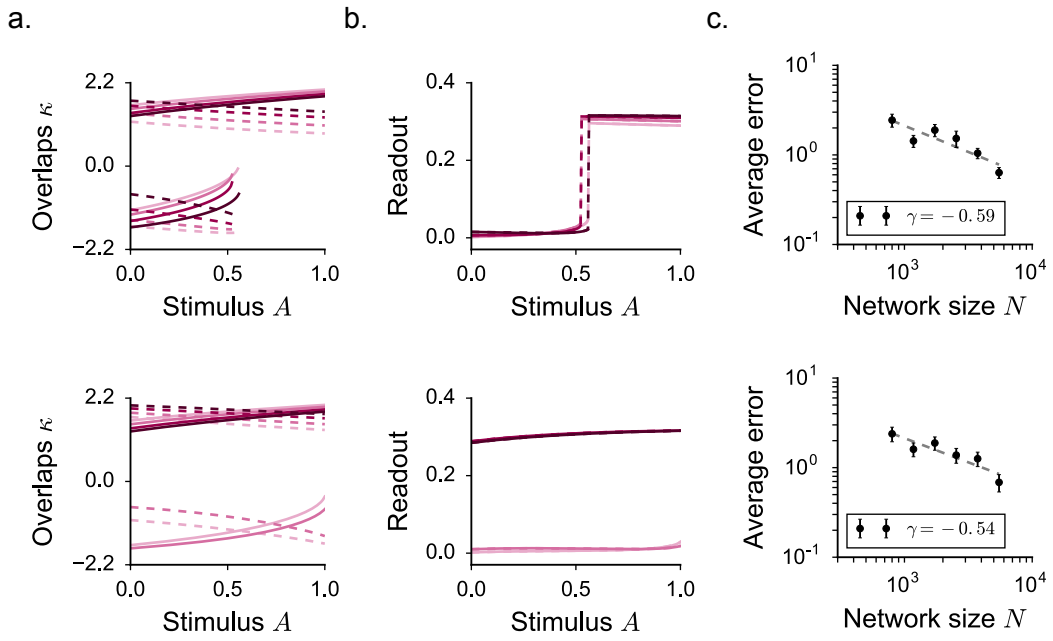


FIGURE 8.8: Rank two structures for implementing non-linear detection in a context-dependent fashion: theoretical mean field predictions. **a.** Values of the first-order statistics κ_1 (continuous) and κ_2 (dashed) as a function of the overlap strength along the stimulus I_A: mean field prediction. The results are shown for four different values of the overlap strength along the second stimulus I_B. Colors and legend as in Fig. 8.7. Top (resp. bottom): context A (B) is selected. **b.** Readout value, built by summing and averaging the values of κ_1 and κ_2 over the initial conditions (Eq. 8.37). Details as in **a**. **c.** Average normalized error in the two contexts as a function of the network size N. Details as in Fig. 7.4 **b**. Parameters as in Fig. 8.7.

from which the value of $\kappa_1 = \kappa_2 = \bar{\kappa}$ can be derived by dividing and multipyling together the two equations. The final readout signal contains a contribution from both first-order statistics:

$$z(t) = \langle w_i[\phi_i] \rangle = \beta_m^2 \Sigma_w^2 (\kappa_1 + \kappa_2) \langle [\phi_i'] \rangle. \tag{8.37}$$

In the present case, the modulatory inputs along I_{ctxA} and I_{ctxB} are used to gate a context-dependent response. Similarly to Fig. 8.2 **c**, a strong and negative gating variable along I_{ctxA} can completely suppress the response to stimulus I_A, so that the readout signal is left free to respond to I_B. Fig. 8.8 **a-b** displays the values of the first-order statistics and the readout response in the two contexts. Note that, when the response to I_A (resp. I_B) is blocked at the level of the readout, the relative first-order statistics κ_1 (resp. κ_2) does not vanish, but it actively contributes to the final network response.

The exact effect of the modulatory inputs is quantified by solving the mean field equations, that can be derived by straightforwardly extending the calculations performed in the unit rank case. For the first-order statistics, we obtain:

$$\begin{aligned} \kappa_1 &= \langle [\phi_i'] \rangle \{ \rho_m \rho_n \kappa_1 + \beta_m \beta_n (\kappa_1 + \kappa_2) + \Delta_A + \rho_n \gamma_A \} \\ \kappa_2 &= \langle [\phi_i'] \rangle \{ \rho_m \rho_n \kappa_2 + \beta_m \beta_n (\kappa_1 + \kappa_2) + \Delta_B + \rho_n \gamma_B \} \end{aligned} \tag{8.38}$$

while the second-order gives, in the case of stationary regimes:

$$\Delta_0 = g^2\langle[\phi_i^2]\rangle + \Sigma_w^2(\kappa_1^2+\kappa_2^2) + \beta_m^2(\kappa_1^2+\kappa_2^2) + 2\Sigma_I^2 + (\rho_m\kappa_1+\gamma_A)^2 + (\rho_m\kappa_2+\gamma_B)^2. \quad (8.39)$$

The average activation variable of single neurons contains entangled contributions from the main directions of the dynamics, which are inherited both from the external inputs and the recurrent architecture:

$$\begin{aligned}\mu_i = [x_i] =& (y_i^{(A)} + \rho_m I_{ctxA,i} + \beta_m w_i)\kappa_1 + (y_i^{(B)} + \rho_m I_{ctxB,i} + \beta_m w_i)\kappa_2 \\ &+ \frac{\Delta_A}{\Sigma_I^2} I_{A,i} + \frac{\Delta_B}{\Sigma_I^2} I_{B,i} + \gamma_1 I_{ctxA,i} + \gamma_2 I_{ctxB,i}.\end{aligned} \quad (8.40)$$

In Fig. 8.7 **b**, we project the average activity $[\phi_i]$ in the directions that are more salient to the task. The projection along w, which reflects the output decision, is proportional to the readout value (Eq. 8.37). The input signals affect instead the average activity through the values of κ_1 and κ_2, but can be also readout directly along the input directions, yielding:

$$\begin{aligned}\langle I_A[\phi_i]\rangle &= \Delta_A\langle[\phi_i']\rangle \\ \langle I_B[\phi_i]\rangle &= \Delta_B\langle[\phi_i']\rangle.\end{aligned} \quad (8.41)$$

Note that the projection on the input direction I_A (resp. I_B) is proportional to the signal Δ_A (resp. Δ_B) regardless of the configuration of the modulatory inputs selecting one input channel or the other.

In more practical terms, in order to obtain the network architecture that has been used in Fig. 8.7, we fixed the parameters step by step. We first considered input patterns only along I_A ($\Delta_B = 0$), and we fix two arbitrary values of β_m and β_n. In particular, we consider intermediate values of β. Large values of β tends to return large variance activity, which require to evaluate with very high precision the Gaussian integrals which are present in the mean field equations. Small values of β bring instead the network activity closer to a continuous-attractor structure, and turn into larger finite-size effects. In a second step, we fix ρ_m and ρ_n such that the network detects normalized input components along I_A only when they are larger than a threshold value, that is taken around 0.5. We then looked for a pair of gating variables strengths $[\gamma_A, \gamma_B]$ which completely suppresses the response to I_A by extending the extent of bistable activity. The opposite pattern can be used to block the response in I_B and allow a response in I_A.

Once the response in I_A has been blocked, it can be verified that the network solely responds to inputs which contain a response along I_B that is larger than a threshold close to 0.5. Note that, similarly to Fig. 8.8 **b**, different values of Δ_A only minimally affect the exact position of the threshold.

To conclude, we remark that this procedure leaves the freedom of fixing the network parameters in many different configurations. The parameters that have been used in Fig. 8.7 have been indicated in the caption. Such complex architecture leads to larger finite-size effects that the respective unit-rank setup which acts as a single detector of correlations. In particular, the error than is performed at the level of the readout is larger but it decays with the system size, as expected for deviations induced by finite-size effects (Fig. 8.8 **c**).

8.6 Oscillations and temporal outputs

We now turn to a final example of dynamical and computational regimes in a network with rank two connectivity structure. We consider a geometrical configuration in which the right

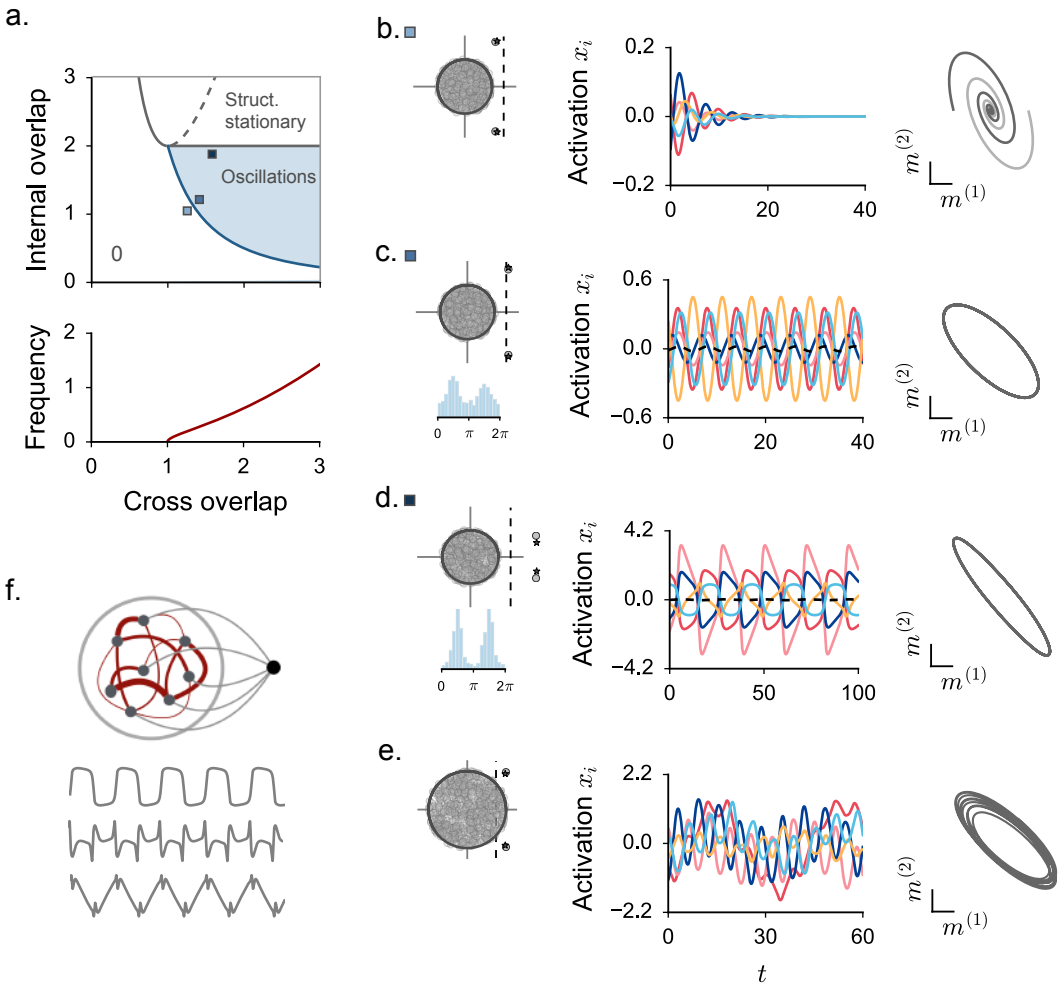
and left vectors corresponding to the two parts of the rank two connectivity structure exhibit cross-overlaps, such that $m^{(1)}$ has a non-zero overlap with $n^{(2)}$, and similary for $m^{(2)}$ and $n^{(1)}$ (see further in Section 8.6.1). We moreover assume that one of these cross-overlaps, e.g. between $m^{(1)}$ and $n^{(2)}$ is negative, so that the two vectors are anti-correlated. In such a configuration, the activity generated along $m^{(1)}$ by the first unit-rank structure will be fed negatively into the second unit-rank structure, giving rise to an effective negative feedback loop. In addition, we assume that some internal overlap is also present between left and right vectors that correspond to the same part of the connectivity structure, e.g. $m^{(1)}$ and $n^{(1)}$ (see Section 8.6.1).

The negative feedback loop implemented by this rank two structure will tend to generate oscillatory activity. Mathematically, this oscillatory activity corresponds to a pair of complex conjugate eigenvalues that lie outside of the continuous part of the spectrum (Fig. 8.9). For moderate amounts of cross-overlap, these eigenvalues do not destabilize the equilibrium activity, but lead to oscillatory transients (Fig. 8.9 **b**). At the level of individual neurons, these transients are highly heterogeneous and multi-phasic, the precise trajectories being determined by the specific set of initial conditions. At the population level, the transients are however dominantly two-dimensional, and lie mainly within the plane defined by the two right-structure vectors $m^{(1)}$ and $m^{(2)}$ as expected from Eq. 8.10. Different initial conditions give rise to different trajectories in the $m^{(1)} - m^{(2)}$ plane that display rotational activity in the same direction. Overall, this rotational transient activity bears a strong resemblance with the population activity recorded during movement onset in the motor cortex [36]. Note that within our framework, the directions capturing rotational activity in the population space can be directly predicted by the connectivity: they are simply given by the right-structure vectors $m^{(1)}$ and $m^{(2)}$.

As the cross-coupling is increased, the complex conjugate eigenvalue cross the stability boundary and give rise to sustained oscillations (Fig. 8.9 **c**). In this dynamical state, the activity of each unit oscillates periodically, but the amplitudes and phases of different units are highly heterogeneous. As a result, different units are out of phase, and the oscillatory activity is not apparent at the population-average level, yet the the population as a whole exhibits rotational activity in the $m^{(1)} - m^{(2)}$ plane. Note that the distribution of phases across units can be directly read from the shape of the trajectories in $m^{(1)} - m^{(2)}$ plane (see Section 8.6.1): symmetric, circular trajectories correspond to a flat distribution of phases, while more elongated trajectories correspond to peaked distributions.

The amount of cross-overlap between left- and right- structure vectors directly controls the frequency of population activity at the oscillation onset (Fig. 8.9 **a**). In contrast, as the internal overlap is increased, the oscillatory activity becomes increasingly non-linear (Fig. 8.9 **d**) and eventually disappears in favor to a dynamical regime in which two equilibrium states are stable (Fig. 8.9 **a**). Close to that transition, highly non-linear oscillatory activity can be understood as periodic jumps between two equilibrium states. On the other hand, increasing the strength of random connections leads to an increase of activity in the directions perpendicular to $m^{(1)}$ and $m^{(2)}$ and eventually generates chaotic activity. As shown for other types of rank one and two connectivity structures, when the strengths of the structured and random parts of the connectivity are comparable, a hybrid regime appears in which the activity displays low-dimensional, structured chaos, in particular quasi-periodic states that resemble mixtures of oscillatory and chaotic activity (Fig. 8.9 **e**). In that regime, numerical simulations show strong finite-size effects that remain to be more fully understood.

The highly heterogeneous oscillatory activity generated by this type of rank two connectivity has interesting computational properties. Since different units have very diverse temporal



profiles of activity, a linear readout unit added to the network can exploit them as a rich basis set for constructing a range of periodic outputs (Fig. 8.9 **f**). A rank-two connectivity structure can therefore be exploited to generate outputs similar for instance to FORCE learning [132] (see Chapter 5).

8.6.1 Mean field equations

We considered the following configuration:

$$\begin{aligned} m^{(1)} &= \alpha x_1 + \rho y_1 \\ m^{(2)} &= \alpha x_2 + \rho y_2 \\ n^{(1)} &= \alpha x_3 + \rho y_2 + \gamma \rho y_1 \\ n^{(2)} &= \alpha x_4 - \rho y_1, \end{aligned} \tag{8.42}$$

where the right- and the left-structure vectors share two cross-overlap directions y_1 and y_2. Note that the vectors in one of the two pairs, $m^{(1)} - n^{(2)}$, are negatively correlated. A second

Figure 8.9 *(previous page)*: Oscillatory activity from rank two structures that include a cross overlap between left- and right- vectors. **a.** Top: phase diagram for the rank two structure we adopt (see *Methods*). For different values of the internal and the cross overlaps, the trivial fixed point can lose stability and give rise to oscillatory or stationary structured activity. The Hopf bifurcation is indicated in blue, the instability to stationary activity in grey. The light-blue parameter region corresponds to sustained non-linear oscillations. Bottom: frequency of oscillations along the Hopf bifurcation boundary, in units defined by the implicit time scale of the network dynamics. **b-c-d-e.** Samples of activity for different connectivity parameters. From left to right: stability eigenspectrum of the trivial fixed point (theory and simulations), sample of activation trajectories (the population average is indicated in dashed black), and population dynamics projected on the right-structure vectors $m^{(1)}$ and $m^{(2)}$. The parameters that have been used for every sample are indicated in **a**. **b**: Oscillatory transients in the fixed point regime. **c**: Stable oscillations above the Hopf instability. The elongated shape of the closed trajectory on the $m^{(1)} - m^{(2)}$ plane is inherited by the phase distribution across the population, and can be tuned by slightly modifying the parameters of the rank two structure (see Section 8.6.1). **d**: Highly non-linear oscillations close to the boundary with bistable activity. **e**: Oscillatory activity at high g values, where dynamics include a chaotic component. **f.** When oscillations are strongly non-linear, their spectrum includes a large variety of frequencies that can be used to reproduce highly non-linear periodic patterns. We designed three random readout vectors and we linearly decoded activity from the dynamical regime in **d** to generate periodic non-linear outputs, which are displayed in grey.

overlap is introduced internally to the $m^{(1)} - n^{(1)}$ pair, and scales with the parameter γ. The directions x_j, with $k = 1, ..., 4$, represent uncorrelated terms. Note that different values of α affect quantitatively the network statistics, but they do not change the phase diagram in Fig. 8.9 **a**.

By rotating P_{ij} on a proper orthonormal basis, one can check that its eigenvalues are given by:

$$\lambda_{\pm} = \frac{\gamma\rho^2}{2}\left(1 \pm \sqrt{1 - \frac{4}{\gamma^2}}\right), \tag{8.43}$$

and they are complex conjugate for $\gamma < 2$. In this case, the internal overlap γ have the effect of returning a non-vanishing real part. The two complex conjugate eigenvectors are given by:

$$e^{\pm} = \left(-\frac{\gamma}{2}m^{(1)} + m^{(2)}\right) \pm i\sqrt{\left|1 - \frac{4}{\gamma^2}\right|}m^{(1)}. \tag{8.44}$$

The eigenspectrum of $J_{ij} = g\chi_{ij} + P_{ij}$ inherits the pair of non-zero eigenvalues of P_{ij}. When $g < 1$ and $\gamma < 2$, the trivial fixed point thus undergoes an Hopf bifurcation when the real part of λ crosses unity (Fig. 8.9 **a**, blue). When $\gamma > 2$, instead, the two eigenvalues are real. One bifurcation to bistable stationary activity occurs when the largest eigenvalue λ_+ crosses unity (Fig. 8.9 **a**, gray).

On the boundary corresponding to the Hopf bifurcation, the frequency of instability ω_H is determined by the imaginary part of Eq. 8.43. At the instability, the oscillatory activity of unit i can be represented as a point on the complex plane. Since close to the bifurcation we can write:

$$\mu_i = e_i^+ e^{i\omega_H t} + c.c.\,, \tag{8.45}$$

its coordinates are given by the real and the imaginary part of the ith component of the complex eigenvector e^+. The phase of oscillation can then be computed as the angle defined by this point with respect to the real axis. Note that the disorder in the elements of the eigenvector e^+, which is inherited by the random distribution of the entries of the structure vectors $m^{(1)}$ and $m^{(2)}$, tends to favour a broad distribution of phases across the population.

In the limit case where the real and the imaginary parts of the complex amplitude of the oscillators are randomly and independently distributed, the population response resembles a circular cloud in the complex plane. In this case, the phase distribution across the population is flat. Note that a completely flat phases distribution can be obtained for arbitrary frequency values by adopting a rank two structure where an internal overlap of magnitude $\gamma\rho^2$ exists between vectors $m^{(2)}$ and $n^{(2)}$ as well.

In the present case, for every finite value of γ, the real and the imaginary part of e_i^+ are anti-correlated through $m^{(1)}$ (Eq. 8.44). Correlations tend to align the network response on two main and opposite phases, as shown in the phase histograms of Fig. 8.9 **c-d**. The distribution of phases becomes sharper and sharper in the $\gamma \to 2$ limit, as the distribution in the complex plane collapses on the real axis.

The phase distribution across the population is reflected in the shape of the closed orbit defined by activity on the $m^{(1)} - m^{(2)}$ plane, whose components are given by κ_1 and κ_2. Because of Eq. 8.10, the phase of the oscillations in κ_1 (resp. κ_2) can be computed by projecting the eigenvector e^+ on the right-structure vectors $n^{(1)}$ and $n^{(2)}$:

$$\begin{aligned} \kappa_1 &= |\kappa_1| e^{i(\Phi_1 + \omega_H t)} + c.c. = \langle n_i^{(1)} [\phi_i] \rangle \\ \kappa_2 &= |\kappa_2| e^{i(\Phi_2 + \omega_H t)} + c.c. = \langle n_i^{(2)} [\phi_i] \rangle \end{aligned} \tag{8.46}$$

By using Eqs. 8.44 and 8.45 we get, in the linear regime:

$$\begin{aligned} \kappa_1 &= \left[\langle n_i^{(1)} m_i^{(2)} \rangle - \frac{\gamma}{2} \langle n_i^{(1)} m_i^{(1)} \rangle + i \langle n_i^{(1)} m_i^{(1)} \rangle \sqrt{\left| 1 - \frac{4}{\gamma^2} \right|} \right] e^{i\omega_H t} + c.c. \\ &= \left[\rho^2 \left(1 - \frac{\gamma^2}{2} \right) + i\gamma\rho^2 \sqrt{\left| 1 - \frac{4}{\gamma^2} \right|} \right] e^{i\omega_H t} + c.c. \end{aligned} \tag{8.47}$$

while:

$$\begin{aligned} \kappa_2 &= \left[\langle n_i^{(2)} m_i^{(2)} \rangle - \frac{\gamma}{2} \langle n_i^{(2)} m_i^{(1)} \rangle + i \langle n_i^{(2)} m_i^{(1)} \rangle \sqrt{\left| 1 - \frac{4}{\gamma^2} \right|} \right] e^{i\omega_H t} + c.c. \\ &= \left[\rho^2 \frac{\gamma}{2} - i\rho^2 \sqrt{\left| 1 - \frac{4}{\gamma^2} \right|} \right] e^{i\omega_H t} + c.c. \end{aligned} \tag{8.48}$$

When γ is close to 2, the complex amplitudes of κ_1 and κ_2 vanish. However, their real part have different sign. We thus get: $\Phi_2 = 0$, $\Phi_1 = \pi$. As a consequence, at large γ values, the oscillatory activity in κ_1 and κ_2 tends to be strongly in anti-phase.

Stationary solutions can be instead easily analyzed with the standard mean field approach. The equations for the first order statistics read:

$$\begin{aligned} \kappa^1 &= (\gamma\rho^2\kappa^1 + \rho^2\kappa^2) \langle [\phi_i'] \rangle \\ \kappa^2 &= -\rho^2 \kappa^1 \langle [\phi_i'] \rangle. \end{aligned} \tag{8.49}$$

The two equations can be combined together to give the following condition on $\langle[\phi_i']\rangle$, which in turn determines the value of Δ_0:

$$\rho^4\langle[\phi_i']\rangle^2 - \gamma\rho^2\langle[\phi_i']\rangle + 1 = 0. \tag{8.50}$$

The mean field equations thus admit two solutions, given by:

$$\langle[\phi_i']\rangle_\pm = \frac{\gamma}{2\rho^2}\left(1 + \pm\sqrt{1 - \frac{4}{\gamma^2}}\right) \tag{8.51}$$

which, similarly to Eq. 8.43, take real values for $\gamma > 2$. Because of the constraints on the sigmoidal activation function, the mean field solutions are acceptable only if $|\langle[\phi_i']\rangle| < 1$. As it can be easily checked, the condition $\langle[\phi_i']\rangle_- < 1$ coincides with imposing $\lambda_+ > 1$. We conclude that two stationary solutions exist above the instability boundary of the trivial fixed point (Fig. 8.9 **a**, gray). A second pair of solutions appears for $\langle[\phi_i']\rangle_- < 1$, which coincide with $\lambda_- > 1$ (Fig. 8.9 **a**, dashed), where the second outlier of J_{ij} becomes unstable. This second pair of solutions is however always dynamically unstable, as it can be checked by evaluating the outliers of their stability matrix through Eq. 46. The coefficients of the reduced matrix $\mathcal{M}$ read:

$$\begin{aligned} a_{11} &= \gamma\rho^2\langle[\phi_i']\rangle \\ a_{12} &= \rho^2\langle[\phi_i']\rangle \\ b_1 &= \frac{1}{2}\rho^2(\kappa^{20} + \gamma\kappa^{10})\langle[\phi_i'']\rangle \end{aligned} \tag{8.52}$$

and

$$\begin{aligned} a_{21} &= -\rho^2\langle[\phi']\rangle \\ a_{22} &= 0 \\ b_2 &= -\frac{1}{2}\rho^2\kappa^{10}\langle[\phi'']\rangle. \end{aligned} \tag{8.53}$$

On the phase diagram boundary corresponding to $\gamma = 2$, the stable and the unstable pair of stationary solutions annihilate and disappear. At slightly smaller values of γ ($\gamma \lesssim 2$), the network develops highly non-linear and slow oscillations which can be though as smooth jumps between the two annihilation points (Fig. 8.9 **c-d**).

8.7 Discussion

Motivated by the observation that a variety of approaches for implementing computations in recurrent networks rely on a common type of structure in the connectivity, we studied a class of network models in which the connectivity matrix consists of a sum of a fixed, low-rank term and a random, full rank part. We found that in this model class, both spontaneous and stimulus-evoked activity in large networks could be described in detail using a mean field analysis. This approach led us to a simple, geometrical understanding of the relationship between connectivity and dynamics, and allowed us to design minimal connectivity structures that implemented specific computations.

Our central result is that the low-rank structure in the connectivity matrix induces low-dimensional dynamics in the network, a hallmark of population activity recorded in behaving

animals [51]. While low-dimensional activity is usually detected using dimensional-reduction techniques [41], our analysis allows to directly identify the low-dimensional subspace that contains the dominant part of the dynamics, and provides a direct interpretation of this subspace in terms of connectivity structure. This relationship between connectivity and dynamics is however highly non-linear, and we have showed that the dynamical repertoire of the network increases quickly with the rank of the connectivity structure. As a consequence, rank two connectivity already leads to a rich range of dynamics that is sufficient to implement complex computations.

A key component of our analysis is the simple fact that a matrix of rank r is fully specified by $2r$ N-dimensional vectors, where N is the size of the network. We have shown that the dynamics in the network can be intuitively understood from the geometrical arrangement of these $2r$ structure vectors with respect to $N-$dimensional vectors representing the patterns of inputs. We have specifically focused on the case where both structure and input vectors are fixed, but generated from a random distribution. While geometry in dimensions larger than three is generally hard to grasp, dealing with a small number of very high-dimensional random vectors is relatively straightforward as their geometry reduces to stochastic calculus. High-dimensional random vectors moreover have very interesting computational properties that have been pointed out within the framework of so-called hyper-dimensional computing [70]. Low-rank random recurrent networks studied here directly inherit properties such as the ability to easily generalize or bind features, and moreover combine theses generic properties with additional non-linear features such as gating. We have showed that these non-linear features can be exploited in particular to implement context-dependent computations.

A traditional approach for implementing computations in recurrent networks has been to endow them with a clustered [147, 78] or distance-dependent connectivity [17]. Such networks inherently display low-dimensional dynamics similar to our framework [78, 150], as clustered connectivity is in fact a special case of low-rank connectivity. The main difference with the framework studied here is that clustered connectivity is highly ordered, since each neuron belongs to a single cluster and therefore is selective to only one feature (a given stimulus, or a given output). Neurons in clustered networks are therefore highly specialized and display so called pure selectivity. Here instead we have considered random low-rank structures, in which stimuli and outputs are represented in a random, highly distributed manner and individual neurons are typically responsive to several stimuli, outputs, or combinations between stimuli and outputs. Such mixed selectivity is a ubiquitous property of cortical neurons [107, 32], and confers additional computational properties related to the hyper-geometrical framework [15]. While the dynamics lie in a low-dimensional subspace, this subspace is randomly embedded within the $N-$dimensional space. Moreover, while many of the dynamical regimes found in our framework are identical to dynamical regimes in networks with clustered, or distance-dependent connectivity, we have shown that the combination of random and structured connectivity can give rise to novel regimes, in which the activity is chaotic, but explores an underlying structure. Such a combination of fluctuating and structured activity can in particular give rise to slow timescales in the dynamics (see Appendix D). Our analyses also show that some dynamical regimes require less connectivity than previously thought. For instance, classical implementations of ring attractors rely on distance-dependent connectivity with a ring structure. Here instead we found that this ring-structure in the connectivity is not necessary, as a rank-two connectivity (with an inherent symmetry) is sufficient to generate ring attractors.

This study is closely related to the classical framework of Hopfield networks [65]. The aim

of Hopfield networks is to store in memory specific patterns of activity by creating for each pattern a corresponding fixed-point in the network dynamics. This is achieved by adding a rank-one term for each memorized item, and one approach for investigating the capacity of such a setup has been the mean field theory of a network with a connectivity that consists of a sum of a rank one term and a random matrix [139, 110, 122]. While this approach is clearly close to the one adopted in the present study, there are important differences. Within Hopfield networks, the unit rank terms are symmetric, so that the corresponding left- and right-structure vectors are identical for each pattern. Moreover, the unit rank terms corresponding to different patterns are generally uncorrelated. In contrast, here we have considered the more general case where the left- and right-eigenvectors are different, and potentially correlated between different rank one terms. Most importantly, our main focus was on responses to external inputs, and input-output computations rather than fixed points of spontaneous activity. In particular we showed that left- and right- structure vectors play different roles with respect to processing inputs, the left-structure vector playing the role of input-selection, and the right-structure vector determining the output of the network. While in Hopfield networks the number of fixed points increases linearly with the rank of the connectivity matrix [10, 2], we have shown that the full dynamical repertoire of the network depends on the geometrical arrangement between left- and right-vectors, the combinatorics of which potentially increases exponentially with the rank of the structure term. Whether the relationship between the dynamical repertoire and the rank of perturbations is really exponential remains to be determined, but our study shows that for rank-two perturbations the dynamics are already very rich and sufficient to implement complex computations.

Our study is also closely related to echo-state networks (ESN) [67, 68] and FORCE learning [132]. In those frameworks, randomly connected recurrent networks are trained to produce specified outputs using a feedback loop from the readout unit to the network. Mathematically, adding a feedback loop is exactly equivalent to adding a rank-one term to the random connectivity matrix [80], where the left-structure vector corresponds to the read-out vector and the right-structure vector corresponds to the feedback (note that in the computational implementations presented here, the readout is instead determined by the right-structure vector). In their most basic implementation, both echo-state and Force learning train only the readout weights, but the details of the learning procedure differ between the two. The training is moreover performed for a fixed realization of the random connectivity, hence the final rank-one structure obtained from Echo-state and Force learning is correlated with the specific realization of the random part of the connectivity. Moreover, the unit rank perturbation may be strong. In contrast, here we studied the situation where the low-rank structure is weak and independent from the random part. How important are the correlations? The answer appears to depend on the specific learning procedure. In Chapter 9, we extend our approach to the specific case of echo-state networks trained to produce a constant output [108]. We show that in the solution found by ESN, the correlations between the rank one structure and the realization of the random matrix are weak and merely act to reduce the error in the readout. This error scales as $1/\sqrt{N}$ in absence of correlations, but is reduced to zero thanks to the correlations between the unit rank term and the random connectivity.

The network model used in this study is the one used in most studies based on trained recurrent networks [132, 73, 15, 100, 45, 83, 134]. While this model is very popular, it is highly simplified and lacks many biophysical constraints, the most basic ones being positive firing rates, the segregation between excitation and inhibition and interactions through spikes. Recent works have investigated extensions of the abstract model used here to networks with

biophysical constraints [58, 69, 87]. The main difficulty is that additional constraints make the mean-field analysis more complex. In the Appendix C, we extend our analysis to positive input-output functions and show that little changes. In a first approximation, excitation-inhibition segregation corresponds to adding an additional (clustered) unit rank structure [87]. Interactions through spikes can be approximated using an additional external noise [87, 55, 69]. Additional investigations will be needed to determine the specific effect on each of these constraints on the dynamics and computations in networks with low-rank connectivity structure.

Despite its highly simplified and abstracted nature, the network model examined here captures and connects a number of outstanding experimental observations. First, as pointed out above, the representations of stimuli and outputs are high-dimensional, distributed and mixed, while the computations are based on low-dimensional dynamics on these representations. Both of these properties are shared by a large number of population recordings in behaving animals [51]. Second, the network naturally reproduces the experimental fact that stimulus onset reduces the variability in neural activity, a property shared by a large number of cortical areas [35]. In our model, this reduction of variability is based on two mechanisms that so far have been considered separately: a reduction of multi-stable activity [78, 44] and a quench of chaotic fluctuations in the network [99]. Third, the unit rank structure inferred from computational constraints reproduces known properties of synaptic connectivity. We have shown that in order to produce desired computations, the left-structure vector needs to be correlated with the pattern of inputs that corresponds to the preferred stimulus. As a consequence, if two neurons both strongly encode that stimulus, their reciprocal connections will be stronger than average. This property directly corresponds to the experimental finding that neurons with similar tuning properties are connected by strong reciprocal synapses [71, 72]. Another computational constraint in our framework is that the right-structure vector is correlated with the output readout vector. This implies that two neurons with strong selectivity for a given decision are also connected by strong recurrent connections. To our knowledge, this prediction remains to be tested experimentally.

The connectivity matrices we considered consisted of an explicit sum of a low-rank and a random part. While this may seem as a severe restriction, in fact any arbitrary matrix can be approximated with a low-rank one by keeping a small number of dominant singular values and associated vectors – the basic principle underlying dimensionality reduction. From this point of view, our theory suggests a simple principle: the low-dimensional structure in connectivity determines low-dimensional dynamics and computational properties of recurrent networks. While more work is needed to establish under which precise conditions this statement holds, this principle provides a simple and practically useful working hypothesis for relating connectivity, dynamics and computations in trained neural networks, and one day in experimental setups in which both activity and connectivity are recorded.

A supervised training perspective 9

Throughout Chapter 8, we have been exploiting the theoretical framework we developed in Chapter 7 to explicitly design robust computational models. As discussed in Chapter 5, a different and widely adopted approach for constructing network models which can operate as computational units consists of supervised training procedures. In those frameworks, the exact output of the network is fixed, and weights updates are applied to the synaptic connections of the network in order to match the network output to the desired target.

In the spirit of helping clarifying the theoretical underpinning of trained network dynamics (see also Chapter 6), here we more directly apply our theoretical framework to such approaches. To this end, we focus on two simple tasks which can be solved by adopting appropriate weak and uncorrelated unit rank structures.

Following the theoretical framework derived in Section 7.3, we consider the mean field equations which describe the resulting computational setup. Similarly to an algorithm which seeks least-square solutions, the mean field equations can be inverted to derive the statistics of the appropriate low rank structure which appropriately fulfills the task. Once the unit rank structure has been determined, the number and the stability of the mean field solutions can be quantitatively assessed. Critically, we find that this approach, where the required low-dimensional structure is computed by blindly fixing the network output, can result in non-trivial instabilities even in extremely simplified computational setups. Furthermore, we show that the same kind of instabilities are observed in traditionally trained networks.

9.1 Input-output patterns associations

In a first step, we consider the network architecture that we extensively analyzed in Chapter 7: a recurrent neural network, whose connectivity structure consists of the sum of a random and a unit rank component, receives a $N-$dimensional pattern $I = \{I_i\}$ as external input. We examine to which extent the pattern of equilibrium firing rates $\phi = \{\phi_i\}$ of individual units can be set to a specified pattern $p = \{p_i\}$ (Fig. 9.1 **a**) by imposing the structure vectors m and n.

As in Chapters 7 and 8, we consider that the disordered part of the connectivity is random and cannot be controlled. We therefore impose the output pattern p only at the level of the

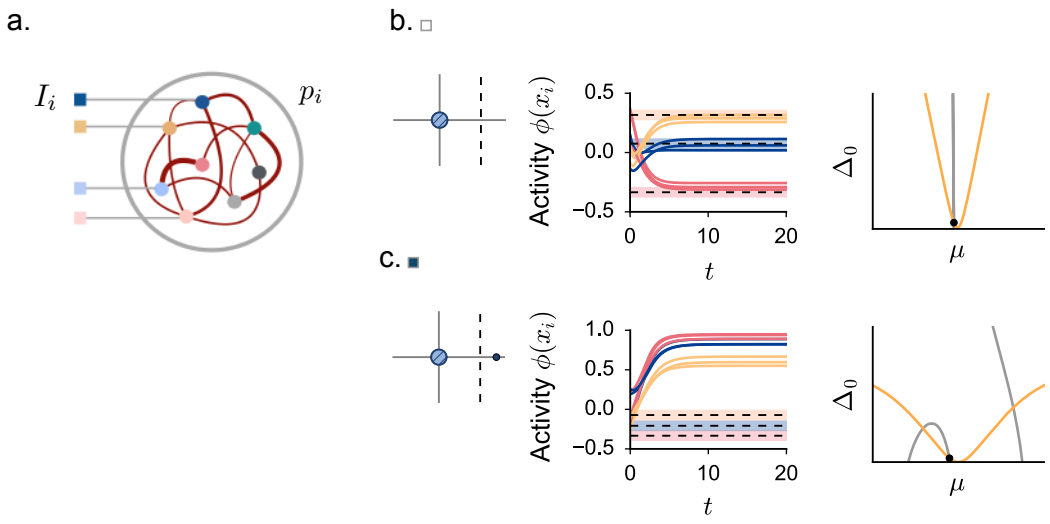


FIGURE 9.1: Implementing specific input-output associations using rank one connectivity structures. **a.** Every unit receives a constant input of value I_i. We aim to construct a rank one connectivity such that the output $\phi(x_i)$ of unit i is on average equal to the target value p_i. **b.** Example of a stable input-output association in a finite-size network. The structure vectors m and n are determined by inverting the DMF equations, as described in the text. The network activity converges close to the imposed output pattern (black dashed lines). The firing rate variables are displayed for three randomly selected units, shown in three different colors. For each unit, temporal trajectories of activity are shown for three different realizations of the random connectivity χ_{ij}. The pattern statistics are indicated by the small squares in the phase diagram of Fig. 9.2 **a**. The side panel shows the stability eigenspectrum as predicted by the mean field analysis, with the vertical black dashed line indicating the stability boundary. The right panel displays the stationary nullclines for the mean field equations of the resulting structured network (see Section 7.3.6). The state corresponding to the desired computation is indicated by the black dot. Note that, together with the desired state, the final network can admit other mean field solutions. **c.** Example of an unstable input-output association. Details as in **b**. Choice of the parameters: see Fig. 9.2.

average with respect to disordered connectivity:

$$[\phi_i] = p_i. \tag{9.1}$$

In the most generic configuration, the input and the target pattern share an overlap direction. With no loss of generality, we focus on the case where the overlap occurs on the unitary direction, which is equivalent to considering input and output vectors of non-vanishing mean.

For a fixed network structure and a given external input, the dynamical mean field theory yields a set of equations for the network output. These equations can in principle be inverted to determine the structure vectors m and n that lead to the desired output p.

9.1.1 Inverting the mean field equations

In Chapter 7, we derived that the direction of the network activation is determined by the input pattern and the right-structure vector m:

$$\mu_i = m_i\kappa + I_i, \tag{9.2}$$

while at the level of the network activity we have:

$$[\phi_i] = \int \mathcal{D}z\, \phi(m_i\kappa + I_i + \sqrt{\Delta_0^I}z). \tag{9.3}$$

Our aim is therefore to solve for m_i and n_i the following equation:

$$p_i = \int \mathcal{D}z\, \phi(m_i\kappa + I_i + \sqrt{\Delta_0^I}z). \tag{9.4}$$

Note that for fixed Δ_0^I, this equation can be rewritten as

$$p_i = \Phi(m_i\kappa + I_i), \tag{9.5}$$

where Φ is a monotonically increasing, sigmoid function obtained by integrating ϕ over the noise term. If κ is moreover known, and non-zero, m_i is simply given by

$$m_i = \frac{\Phi^{-1}(p_i) - I_i}{\kappa}. \tag{9.6}$$

In the final configuration, the direction of m is then determined by the geometrical arrangement of vectors p and I. The main task is to determine the network-averaged quantities κ and Δ_0^I which are consistent with the specific input-output transformation.

As noted in the previous Chapters, the left- and right-structure vectors m and n play different roles in producing the output pattern. The elements of the vector m directly determine the output of the network, but this output is realized only if the overlap κ is non-zero. The role of n is precisely to yield a non-zero overlap κ with the network activity by selecting appropriate inputs. As described in Section 7.2, a non-zero overlap can be obtained in several manners, in either by having a non-zero overlap between n and m or between n and I, or a combination of both. In Section 8.1, non-vanishing values of κ have been obtained by tuning n along the component of the input vector that is perpendicular to m. Here, we focus here on a different geometrical arrangement, in which I and n overlap in the unitary direction, which is shared between I and p (and thus m). As it will be shown in Section 9.2.1, this configuration resembles the network structure that emerges from numerical training and results in non-trivial stability properties.

We then take M_I, $M_n > 0$. We moreover set $M_n = 1$ and we indicate by M and Σ the mean and variance of the unknown structure vector m. Eq. 9.4 thus transforms into:

$$p_i = \int \mathcal{D}z\, \phi(m_i\langle[\phi_i]\rangle + I_i + \sqrt{\Delta_0^I}z). \tag{9.7}$$

The value of m_i can thus be computed once the mean field statistics $\langle[\phi_i]\rangle$ and Δ_0^I have been computed. Their value depends self-consistently on the value of the input and the output

patterns that we are trying to impose, so that additional constraints on the activity statistics have to be taken into account.

For simplicity, we focus on a network in a stationary state; similar arguments can be used to derive the equations for the chaotic case. The network activity is determined by the two mean field equations:

$$\begin{aligned}\mu &= M\langle[\phi_i]\rangle + M_I \\ \Delta_0 &= g^2\langle[\phi_i^2]\rangle + \Sigma_\mu^2\end{aligned} \tag{9.8}$$

with $\Sigma_\mu^2 = \Sigma^2\langle[\phi_i]\rangle^2 + 2\Sigma_{mI}\langle[\phi_i]\rangle + \Sigma_I^2$. Eq. 9.8 contains 5 unknown variables: the two activation statistics (μ, Δ_0) and the structure parameters (mean M, variance Σ of the vector m and correlation Σ_{mI} between m and I). In order to close the system, we derive three additional equations by constraining the output activity statistics to be compatible with the output pattern p_i.

We start by imposing:

$$M_p = \langle p_i\rangle = \int \mathcal{D}z\phi(\mu + \Delta_0) = \langle[\phi_i]\rangle. \tag{9.9}$$

Similarly, for the second order statistics, we impose $\langle[\phi_i]^2\rangle = \langle p_i^2\rangle = M_p^2 + \Sigma_p^2$, which reads:

$$\begin{aligned}M_p^2 + \Sigma_p^2 &= \langle\left[\int \mathcal{D}z\ \phi(m_i\langle[\phi_i]\rangle + I_i + \sqrt{\Delta_0^I}z)\right]^2\rangle \\ &= \int \mathcal{D}y\left[\int \mathcal{D}z\ \phi(\mu + \Sigma_\mu y + \sqrt{\Delta_0^I}z)\ \right]^2.\end{aligned} \tag{9.10}$$

The last condition on Σ_{mI} comes instead from imposing:

$$M_I M_p = \langle I_i p_i\rangle = \langle I_i[\phi_i]\rangle = \langle I_i \int \mathcal{D}z\phi(m_i\langle[\phi_i]\rangle + I_i + \sqrt{\Delta_0}z)\rangle. \tag{9.11}$$

Similarly to standard DMF derivations, we explicitly build m_i and I_i as correlated Gaussian variables to get:

$$\begin{aligned}M_p M_I = \int \mathcal{D}w \int \mathcal{D}x_2 (M_I + \sqrt{\Sigma_I^2 - \Sigma_{mI}}x_2 + \sqrt{\Sigma_{mI}}w) \int \mathcal{D}z\ \phi(\mu \\ + \sqrt{\Sigma_{mI}}(1 + \langle[\phi_i]\rangle)w + \sqrt{\Delta_0^I + (\Sigma^2 - \Sigma_{mI})\langle[\phi_i]\rangle^2}z + \sqrt{\Sigma_I^2 - \Sigma_{mI}}x_2).\end{aligned} \tag{9.12}$$

A little algebra, together with Eq. 7.27, returns the final result:

$$\Sigma_{mI} = -\frac{\Sigma_I^2}{M_p}. \tag{9.13}$$

We are thus left with the final set of equations in $(\mu, \Delta_0, M, \Sigma)$:

$$\begin{aligned}&\mu = M\langle[\phi_i]\rangle + M_I \\ &\Delta_0 = g^2\langle[\phi_i^2]\rangle + \Sigma_\mu^2 \\ &M_p = \langle[\phi_i]\rangle \\ &\Sigma_p^2 + M_p^2 = \langle[\phi_i]^2\rangle = \int \mathcal{D}x\left[\int \mathcal{D}z\phi(\mu + \Sigma_\mu x + \sqrt{\Delta_0^I}z)\right]^2\end{aligned} \tag{9.14}$$

with $\Sigma_\mu^2 = \Sigma^2 M_p^2 - \Sigma_I^2$. Once the self-consistent network statistics have been computed as a solution of Eq. 9.14, the structure eigenvector m can be found by solving Eq. 9.7 for every unit i.

Clearly, not every output pattern can be implemented by the network. In particular, for each i, p_i needs to lie in the output range of the sigmoidal transfer function $\phi(x)$ (e.g. for $\phi(x) = \tanh(x)$, we can only expect to correctly reproduce the patterns for which $|p_i| < 1$ for all units). A significant additional complication is that the obtained solution may not be stable with respect to the dynamics of the recurrent network. In order to assess the stability of an obtained solution, we therefore determine the properties of its stability eigenspectrum by computing the value of the radius and the position of the outlier eigenvalues in the eigenspectrum (Eqs. 7.33 and 7.88).

9.1.2 Stable and unstable associations

Fig. 9.1 **b** illustrates a case in which the output pattern p_i is correctly implemented: for any realization of the random part of the connectivity χ_{ij}, the network rate variables converge close to the target pattern. In the stability eigenspectrum predicted by the mean field theory, all the eigenvalues lie well below the instability boundary. A nullcline plot reveals that such fixed point is the only stationary state of the dynamics.

Fig. 9.1 **c** shows instead a case in which the target output corresponds to an unstable solution of the dynamics, and therefore cannot be reached. Iterating the recurrent dynamics with the computed structure vectors m and n, brings the network activity far from the desired output pattern p. The stability eigenspectrum includes an unstable outlier eigenvalue above the instability threshold. The nullcline plot reveals that in this case, the solution found by solving the system in 9.14 corresponds to the unstable fixed point built on the intermediate branch of the μ nullcline (see Section 7.3.8).

More generally, the set of outputs that can be implemented in a stable manner can be determined as function of the statistics of input and output patterns. The corresponding phase diagram (Fig. 9.2 **a**) shows that unstable solutions occupy a sizable portion of the parameter space.

By inverting the mean field equations, we imposed the prescribed output on the level of the average with respect to the random connectivity. In a specific instantiation of the random connectivity, the actual network output will deviate with respect to the average output. The magnitude of the error in the output depends on network parameters and can be directly computed within our theoretical framework. When the output pattern is correctly learnt, our mean field framework allows us to estimate the variability between different realizations of the random connectivity, and therefore the error in the output (Fig. 9.2 **b**). The amplitude of the error at the level of individual units (defined as $C_0^I = \langle [\phi_i^2] \rangle - \langle [\phi_i]^2 \rangle$) depends on the patterns parameters and in general increases super-linearly with the strength g of random connections. For $g > 1$, the network can approximate the pattern p by producing either static or structured chaotic activity, in which the activity of unit i is on average p_i, but temporal fluctuations are present in addition to static ones. In this regime of strong disorder, the error at the level of individual units is large. If the similarity to the target pattern p is however quantified at the population level by computing the normalized overlap between the network activity ϕ and the pattern p, the variability in individual units compensate each other, and the error decreases as $1/\sqrt{N}$ in finite networks (Fig. 9.2 **c**).

To conclude, we have shown that our theoretical approach allows us to determine one-

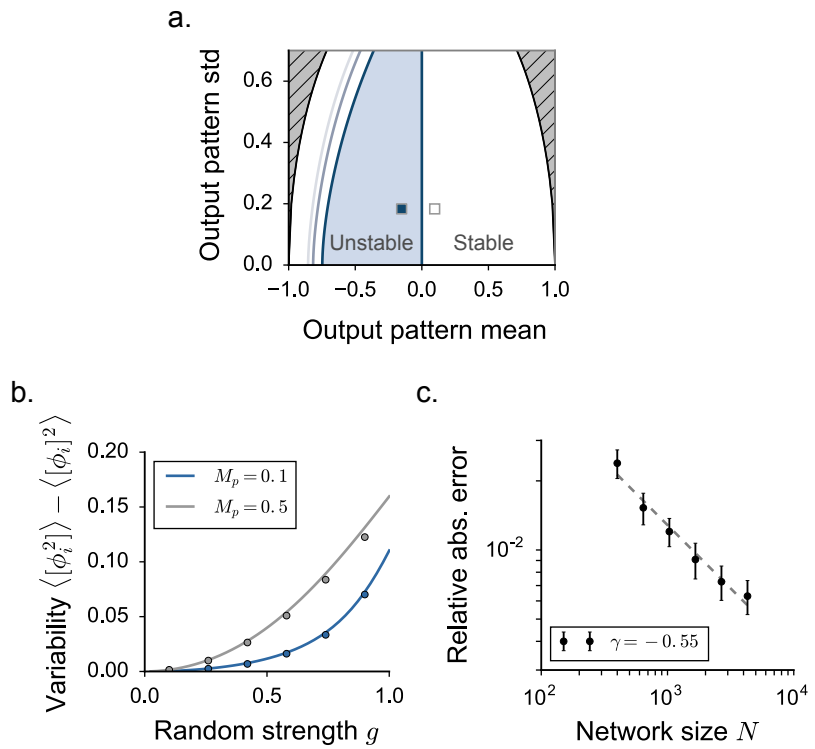


Figure 9.2: Implementing specific input-output associations using rank one connectivity structures: performance. **a.** Stability of the solutions as a function of the output pattern statistics. White (resp. light blue) areas: the association is stable (resp. unstable). Hatched areas: the association cannot be realised as the output statistics are incompatible with the shape of the sigmoidal activation function. The instability region is displayed for three values of the mean input pattern M_I, ranging from 0.8 (dark) to 2 (light). **b.** When the output pattern is correctly learnt, the output activity of each unit fluctuates around the target value p_i. The amplitude of this variability can be measured as $C_0^I = \langle [\phi_i^2] \rangle - \langle [\phi_i] \rangle^2$, and depends on the statistics of the input and the output patterns. Here, the magnitude of the variability is shown for two different values of the mean of the output pattern M_p. Continuous lines: DMF theoretical prediction. Dots: average variability measured for 4 input-output associations. For every pair, C_0^I is measured across 15 realizations of the random connectivity χ_{ij}. **c.** Although single units activity can strongly fluctuate around the output pattern, this variability is largely washed out when the activity vector ϕ is projected on the output p. The value of the projection can be predicted with the mean field theory. In finite size networks, small deviations from the theoretical expectation can be measured because of finite size effects. As expected, the magnitude of the average normalized error decays with the network size. In grey: $g = 0.2$, in black: $g = 1$. Dashed lines: power-law best fit ($y \propto N^\gamma$). The values of γ are indicated in the legend. In this figure, if not differently indicated, $g = 0.2$, $\Sigma_I = 0.1$. The input and the output pattern vectors are uniformly distributed.

dimensional connectivity structures that lead to a specified output in response to a given input and to assess the stability of the resulting recurrent network. We found that unstable associations can be realised for certain values of the statistics of the input and the output patterns. Unstable computations derive from the presence of non-zero asymmetric solutions in the bistable dynamical regimes where the network is projected into (see Section 7.3.8). As in Fig. 7.1, bistable activity can appear because of non-vanishing overlaps between the input, the output and the left-structure vector, that are amplified when the mean field solutions are inverted.

This result suggests that, for the present task, exploiting non-bistable regimes might result in better training. That can be achieved, similarly to Section 8.1, by imposing non-vanishing values of κ through aligning the left-structure vector n solely along the non-shared component of I. In this case, as it has been shown Fig. 7.3 **b-c-d**, a single stable state is created.

Having found that this simple task admits two conceptually alternative solutions, we turn to investigate the strategy that is adopted by simple training procedures. Note that the solutions which are found by our framework are restricted to the class of weak low-dimensional matrices that are not fine-tuned to a specific realization of the noise in the random bulk connectivity matrix. On the contrary, trained networks might implement the same task by exploiting both stronger scaling properties and fine-tuned correlations between the random and the structured connectivity.

In the next section, we address this question in the setup of a a classic echo-state architecture (see Chapter 5). By exploiting our theoretical framework, we design a mean field solution which rely on weak and uncorrelated structures. We then compare our result which the solution that is returned by a least-square batch update and with the dynamics which is directly measured in trained networks.

9.2 Input-output associations in echo-state architectures

In echo-state computing, the network output is not defined for every unit, but only at the level of a single readout $z(t)$. The readout activity is given by a linear combination of the network rate variables:

$$z(t) = \sum_{j=1}^{N} w_j \phi(x_j(t)). \tag{9.15}$$

As the readout signal is fed back to the network, the reservoir dynamics read:

$$\dot{x_i}(t) = -x_i + g \sum_{j=1}^{N} \chi_{ij} \phi(x_j(t)) + u_i z(t) + I_i. \tag{9.16}$$

Combined together, Eqs. 9.15 and 9.16 describe a network architecture which is equivalent to the network model with unit rank structure that we considered so far, with $u_i = m_i$, $w_i = n_i$ and $z = \kappa$.

As in echo-state machines, we consider the feedback weights m to be fixed. Without loss of generality, we fix $m_i = 1 \; \forall i$. We aim at deriving the left-structure vector n which allows the reservoir network to associate an output constant signal: $z(t) = A$ to an external input pattern I (Fig. 9.3 **a**).

We start by deriving a network solution from the usual mean field framework, and we assess its stability.

The mean field equations which describe the desired dynamics ($\kappa = z = A$) are:

$$\begin{aligned} \mu &= A + M_I \\ \Delta_0 &= g^2 \langle [\phi_i^2] \rangle + \Sigma_I^2 \end{aligned} \tag{9.17}$$

if the network is in a stationary state. Eq. 7.66 can be used to trivially extend this reasoning to chaotic states. In terms of the echo-state training procedure (see Chapter 5), the equations in 9.17 coincide with the mean field characterization of the open-loop system, where the feedback signal is clamped to the target A [108]. We first solve Eq. 9.17 to derive the self-consistent network statistics μ and Δ_0. As a second step, similarly to Section 9.1, we design a left-structure vector n which overlaps with m, thus resulting in non-vanishing κ values. We obtain $A = \kappa = M_n \langle [\phi_i] \rangle$, which gives:

$$M_n = \frac{A}{\langle [\phi_i] \rangle} \tag{9.18}$$

and we design n as any random vector of fixed mean M_n. Note that the same choice of n equivalently applies to every realization of the random bulk χ_{ij}.

Once the connectivity structure has been fixed, the mean field equations can be reshaped back into the usual form:

$$\begin{aligned} \mu &= M_n \langle [\phi_i] \rangle + M_I \\ \Delta_0 &= g^2 \langle [\phi_i^2] \rangle + \Sigma_I^2 \end{aligned} \tag{9.19}$$

which coincides with the whole-network (or closed-loop) mean field description. By using the standard DMF tools (see Chapter 7), we can thus evaluate the stability of the solution and the total number of stable states admitted by the network dynamics.

In Fig. 9.3 **b**, we display the transient activity of a network that has been trained to match the target signal starting from a static and a chaotic dynamical regimes. In the two examples, the association is stable and the readout signal $z(t)$ converges to the target A.

In Fig. 9.3 **c**, we fix the target to $A = 1.3$, and look at the readout values predicted by the mean field theory for increasing values of the mean of the external input pattern. Similarly to Section. 9.1, we find that, when the external input has a non-vanishing overlap with the structure vectors ($M_I \neq 0$), the stability of computations is not always guaranteed. In particular, when M_I is large in absolute value, the association is stable and corresponds to the only possible network solution. The association is instead unstable in the range of values: $-A < M_I < M_I^*$, where the target $z = A$ corresponds to the intermediate and unstable branch of the mean field solutions. Finally, for an intermediate range of values above M_I^*, the association is stable, but a second stable state is admitted by the dynamics. Similarly to the previous section, bistability is induced by non-vanishing overlaps between the structure vectors m and n, and can be avoided by tuning n along the direction of I that is orthogonal to the unitary one.

Fig. 9.3 **d** indicates the stability of the input-output association in terms of the readout target A and the mean input M_I. Note that the stability boundaries only display a weak dependence with respect to the random strength g. They also display a dependence on the input pattern Σ_I (not shown): larger Σ_I values have the effect of extending the instability surface.

To conclude, we observe that our theoretical framework allows to exactly match the output z to the target A solely in the case of infinite size networks. In finite networks, small fluctuations of order $\mathcal{O}(1/\sqrt{N})$ are expected to appear at the level of the readout.

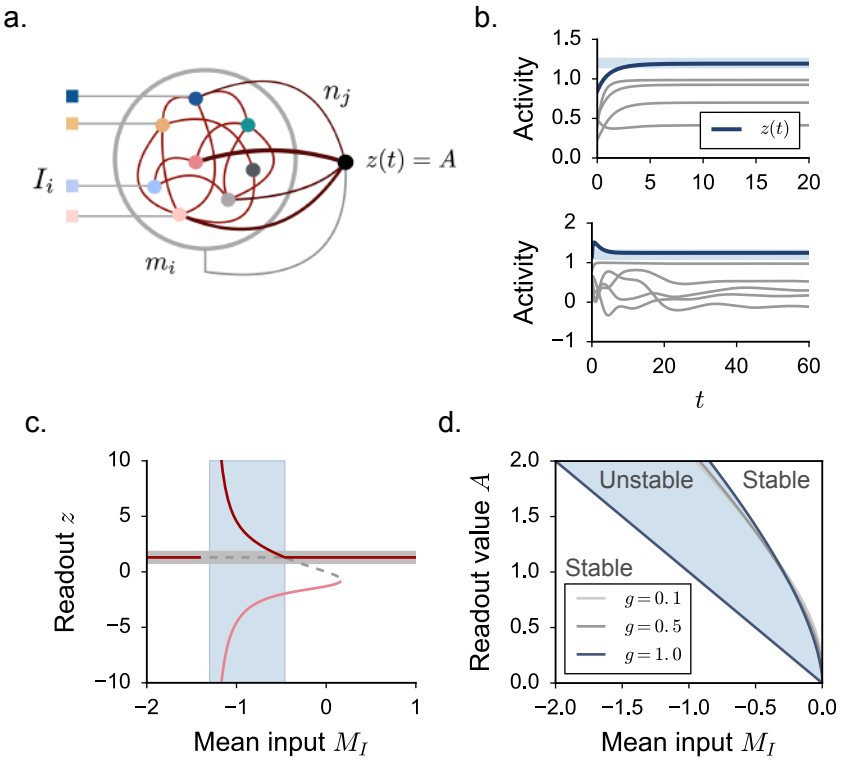


FIGURE 9.3: Designing specific rank one structures which force echo-state architectures to reproduce a constant fixed point at the level of the readout. **a.** The network output is defined at the level of $z = \kappa$, which also acts as a stabilizing feedback input. For a fixed input pattern I, we want the network to respond with a constant signal: $z = A$. **b.** We derive the appropriate structure statistics by solving Eq. 9.17 and we let the network dynamics evolve. Examples from the parameter region where the association is stable. Top: network in a stationary, and down: network in a chaotic regime. Light blue: target, dark blue: network reconstruction, and grey: activity traces for randomly selected units in the network. **c.** From the closed-loop mean field equations 9.19 we predict the value of the readout signal z (grey horizontal line: target A). The two stable solutions are drawn in red, the unstable one in grey. Shaded area: the input-output association corresponds to the unstable mean field solution. Choice of the parameters: $g = 0.5$, $\Sigma_I^2 = 0$. **d.** Stability phase diagram for fixed $\Sigma_I = 0$. The stability boundary weakly depends on the value of g.

9.2.1 A comparison with trained networks

When echo-state networks are numerically trained, finite-size reservoirs are employed. The unit rank structure which is returned by training algorithms is typically fine-tuned to the specific realization of the noise in the random bulk χ_{ij}. By exploiting such correlations, the match between the readout and the target can be made exact for arbitrary network sizes.

When the readout weights n_j are trained offline through least-square error minimization [67], an explicit expression for n_j can be derived [108]. First, the asymptotic open-loop solu-

tion is measured numerically:

$$x_i^* = m_i A + I_i + g \sum_{j=1}^{N} \chi_{ij} \phi(x_j^*) \tag{9.20}$$

and second, the decoding weights are set accordingly:

$$n_j = A \frac{\phi(x_j^*)}{N \langle \phi^2(x_j^*) \rangle}. \tag{9.21}$$

The effect of such a choice is to obtain exactly $z = A$ for whatever network size.

Note that the least-square minimization returns a structure solution whose scaling is weak ($\mathcal{O}(1/N)$) as in our model architecture.

Note also that such training setting can only operate in conditions where the open-loop solution (Eq. 9.20) is stable with respect to chaos. On the contrary, the mean field approach developed in the previous paragraph can be easily extended to take temporal fluctuations into account.

From Eq. 9.20, we expect the open loop activity $\phi(x_j^*)$ to include a component along the right-structure vector m and one along the input I. Since the left vector n is taken to be proportional to $\phi(x_j^*)$ (Eq. 9.21), the training procedure naturally introduces an overlap direction between the two structure vectors m and n. In order to facilitate a direct comparison with our mean field theory, we consider the case where I is parallel to m ($\Sigma_I = 0$), so that the value of κ only includes the effect of the overlaps along the unitary direction.

We then design finite-size echo-state networks by following the offline least-square approach. In Fig. 9.4 **a**, we measure the final value of the readout unit z. Remarkably, we observe that the dynamics of the resulting network matches well the DMF prediction that we derived for rank one structures which are totally uncorrelated with respect to the random bulk of the connectivity χ_{ij} (Fig. 9.3 **c**). More specifically, the mean field predicts with good accuracy the stability of the input-output association and the number of global attractor of the dynamics.

We further look at the stability matrix of the final fixed point, and we find that its eigenspectrum consists of a compact dense component and of a single outlier eigenvalue [108]. In Fig. 9.4 **b**, we show that the position of the outlier is in good agreement with our mean field prediction, which can be computed as in Section 7.3.4.

In [108], it has been shown that when the vector n is set by offline training (Eq. 9.21), the position of the outlier eigenvalue can be exactly computed using control theory arguments. We find that the prediction by [108] almost coincides with the one returned by our mean field theory, although only one branch of the global network solution (the one corresponding to $z = A$) is retrieved in this case.

This strict quantitative agreement suggests that, at least in the present setting, the dynamics of the trained network is only minimally affected by the fine-tuned correlations existing by the finite-sized bulk χ_{ij} and the solution n_j. In particular, the least-square algorithm seems not to be able to exploit such correlations in order to improve the stability of the final trained network.

Similarly to the two cases we studied so far, an instability occurs when the structure overlap induced by the training algorithm becomes strong. In those cases, batch procedures based on the least-square approach are in fact not able to tell apart dynamically stable from unstable

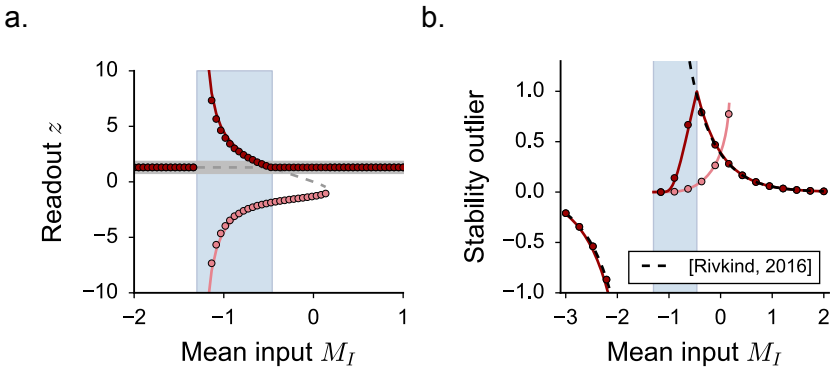


FIGURE 9.4: Echo-state architectures: comparing mean field solutions with the output of trained reservoirs. The training procedure is an offline least-square error minimization. **a.** The mean field prediction (continuous lines), computed as in Fig. 9.3 **c**, predicts with good accuracy the behaviour of trained networks (dots). The network readout converges to A (horizontal gray lines) only outside the instability region (shaded area). **b.** Outlier eigenvalue in the stability eigenspectrum of the final fixed point. Continuous lines: mean field prediction for the three global network solutions. Black dashed lines: position of the outlier which includes the correlations between the bulk and the structure [108]. Dots: eigenvalue position as measured in trained networks. In simulations, choice of the parameters: $g = 0.5$, $\Sigma_I = 0$, $N = 2000$.

solutions. Note that the same kind of instability occur in the more general case where the input vector is taken to be orthogonal to m, because of the non-vanishing correlations between m and n that are naturally induced by Eq. 9.21. As we discussed in Section 9.1, our theoretical framework suggests that a possible strategy to circumvent instabilities consists of operating with non-vanishing κ values that are generated by vectors n and I overlapping on a direction perpendicular to m. Here we found that the least-square solutions adopt instead a non-optimal strategy.

Finally, the least-square approach permits to design a left vector n which allows the network to perform multiple input-output associations with a unique rank one structure. In order to study such setup, the fine-tuned correlations between n_j and χ_{ij} must be crucially taken into account, so that our theory cannot be directly applied. Such setup, however, seem to suffer of even larger stability issues [108], suggesting that also in that case the training procedure is not able of exploiting correlations to improve the stability of the final network.

9.2.2 Different activation functions

To conclude, we present a simple application of our theoretical setup.

In [108], the authors focus on the study of the fixed point task in absence of external inputs. At the end of the analysis, they discuss the dependence of stability properties of the solution on the exact shape of the activation function $\phi(x)$. It is found that adopting a classical sigmoidal function ($\phi(x) = \tanh(x)$) guarantees stability for any target value. On the contrary, the input-output association is always unstable when the activation function is threshold linear with positive offset. Stability depends instead on the architecture parameters when a saturation threshold is imposed on $\phi(x)$.

Here we observe that this result can be easily understood within our framework. In a mean

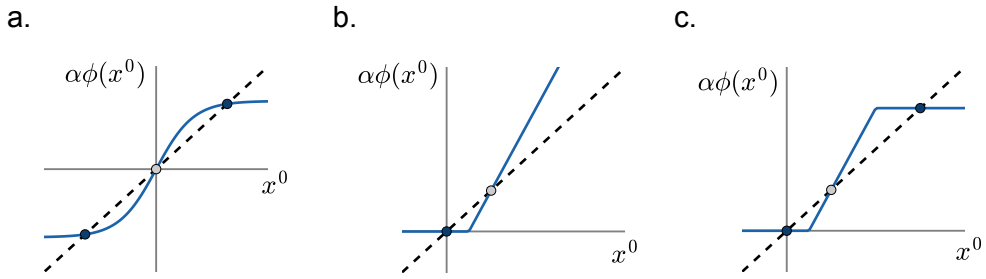


Figure 9.5: Echo-state architectures: dependence of stability on the activation function $\phi(x)$, as in [108]. Graphical solution to Eq. 9.22. Black (grey) dots indicate stable (unstable) solutions. **a.** In the classical case, with $\phi(x) = \tanh(x)$ all the fixed point tasks are stable in absence of external inputs. The only unstable fixed point corresponds indeed with the trivial one. **b.** When the activation function is threshold linear, the only non-trivial fixed point is always unstable, as it is built on the high-gain branch of $\alpha\phi(x)$. **c.** When a saturation value is added, a second stable fixed point appears; if the solution is built on the saturating branch of the activation function, the input-output association is stable.

field perspective, the solution is built by tuning the overlap between the two structure vectors along the unitary direction. The structure of the resulting fixed points can be readout from the noise-free case ($g = 0$), which reads (see Section 7.3.6):

$$x^0 = \alpha\phi(x^0) \tag{9.22}$$

where α measures the structure strength. In Section 7.3.6, we have shown that adding noise in the random bulk and additional directions to the structure vectors does not change the stability properties of the resulting stationary solutions.

In the classical case with $\phi(x) = \tanh(x)$, Eq. 9.22 determines two non-trivial symmetric fixed points (Fig. 9.5 **a**). Because of the sigmoidal saturation, it is easy to check that both non-trivial solutions are stable ($\phi'(x^0) < 0$ in $x^0 \neq 0$). As a consequence, in absence of external inputs, every fixed point is expected to be stable. This result is in agreement with the results from the previous paragraph where we take $M_I = 0$: bistability exists but the only unstable solution is locked in zero, so that it is never selected by training.

When a threshold linear function is adopted, one stable and one unstable fixed points are created. The stable one is locked in $x^0 = 0$, so that non-trivial fixed points are always built on the unstable solution, and training is expected to fail. Finally, adding a saturating threshold induces the presence of a second stable fixed point, so that non-trivial solutions can both be stable or unstable according to the readout target and the architecture parameters. As we saw, both solutions are indeed indistinguishable from the point of view of the training algorithm.

An analysis of network models with more general positively defined activation functions is presented in Appendix C.

Appendix A
Finite size effects and limits of the DMF assumptions

We test numerically the validity of the Gaussian assumptions and the predictions emerging from the DMF theory in the case of excitatory-inhibitory networks (see Chapters 2, 3, 4). We found two main sources of discrepancies between the theory and numerics, namely finite-size effects and the asymmetry between excitation and inhibition.

Finite size effects

As a first step, we analyzed the magnitude of finite size effects deriving from taking finite network sizes. Fig. 1 **a** shows a good agreement between simulated data and theoretical expectations. The magnitude of finite size effects shrinks as the network size is increased and cross-correlations between different units decay (Fig. 1 **c**, left panel).

In Fig. 1 **b** we tested instead the effect of increasing the in-degree C when N is kept fixed. When C is constant and homogeneous in the two populations, our mean field approach requires network sparseness ($C \ll N$). Consistently, we find an increase in the deviations from the theoretical prediction when C is increased (Fig. 1**c**, right panel).

Both the N and C dependencies have the effect of weakly reducing fluctuations variance with respect to the one expected in the thermodynamic limit. The numerically obtained x distribution is in good agreement with the assumption of DMF, which states that current variables x_i are distributed, for large time t and size N, according to a Gaussian distribution of mean μ and variance Δ_0.

Correlations for high ϕ_{max}

We observe that stronger deviations from the theoretical predictions can arise when the upper-bound ϕ_{max} on the transfer function is large and the network is in the intermediate and strong coupling regime. By simulating the network activity in that case, we observe stronger cross-correlations among units, which can cause larger fluctuations in the population-averaged firing rate.

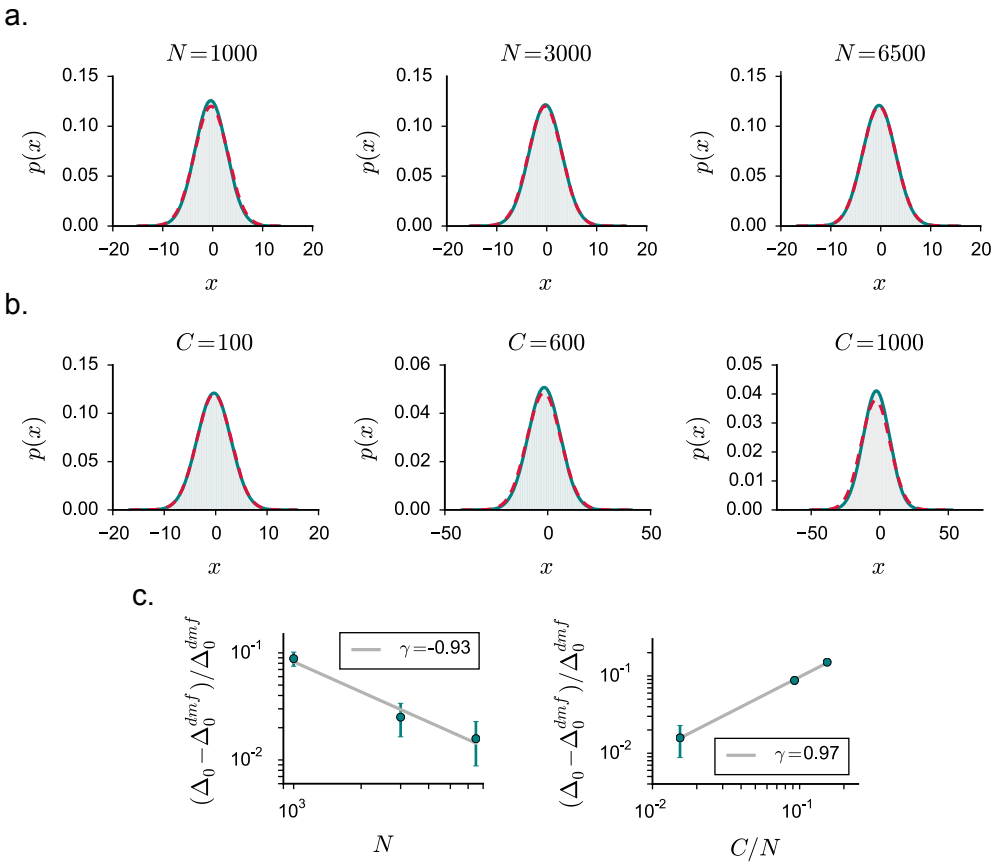


Figure 1: Comparison between dynamical mean field predictions and numerical simulations: general finite size effects. **a.** Dependence on the system size N ($C = 100$). Fistribution of the input current x in the population and in different time steps. The numerical distribution is obtained through averaging over 3 realizations of the synaptic matrix. Light green: simulated data distribution, dark green: best Gaussian fit to data, red: DMF prediction. **b.** As in **a**, dependence on the in-degree C ($N = 6500$). **c.** Normalized deviations from the DMF theoretical value. The log-log dependence is fitted with a linear function, γ giving the coefficient of the linear term. Choice of the parameters: $g = 4.1$, $J = 0.2$, $\phi_{max} = 2$.

In Fig. 2 **a** we check that those deviations can still be understood as finite size effects: the distance between the DMF value and the observed ones, which now is larger, decreases with N as the correlation among units decay. Equivalently, the variance of the fluctuations in the population-averaged input current and firing rate decays consistently as $\sim 1/N$.

The same effect, and even stronger deviations, are observed in rate models where the transfer function is chosen to mimic LIF neurons.

As a side note, we remark that strong correlations in numerical simulations are observed also in the case of spiking networks of LIF neurons with small refractory period and intermediate coupling values (Fig. 2 **b**). Also in this case, correlations are reflected in strong time fluctuations in the population averaged firing rate. Their amplitude should scale with the system size as $1/N$ in the case of independent Poisson processes. This relationship, which is well fitted in the weak and strong coupling regimes (not shown), appears to transform into a weaker

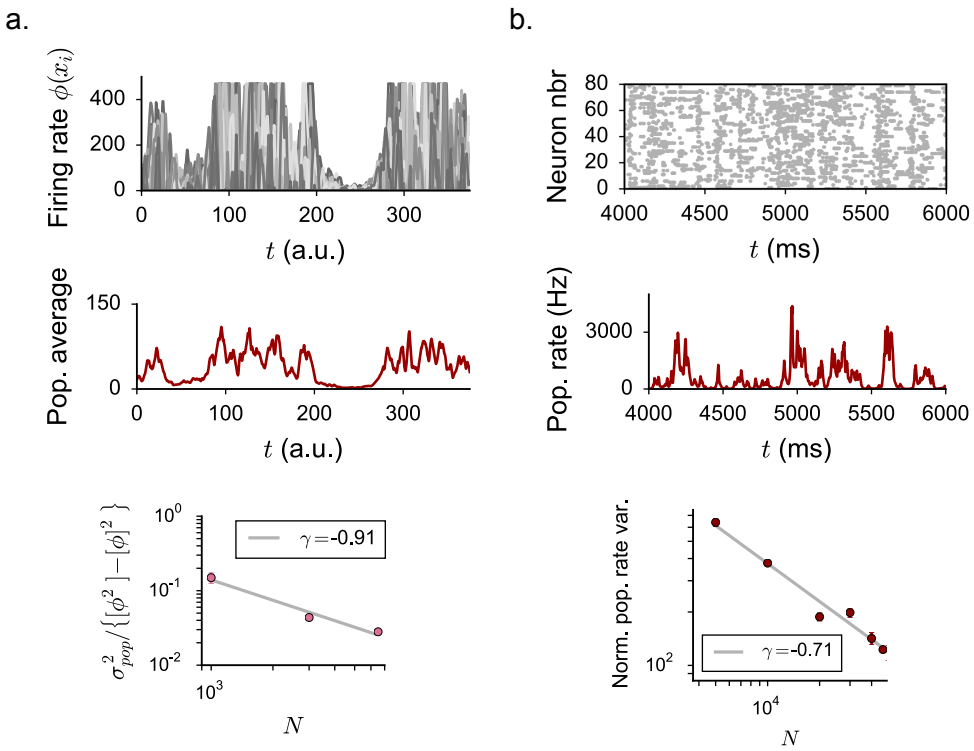


FIGURE 2: Comparison between dynamical mean field predictions and numerical simulations: high saturation bounds. **a.** Finite size effects in rate networks with large saturation upper-bound. Top: sample of network activity, single units. Middle: population averaged firing rate. Choice of the parameters: $g = 5$, $J = 0.14$, $\phi_{max} = 240$. Bottom: normalized variance of the population-averaged firing rate as a function of the network size. **b.** Finite size effects in networks of LIF neurons with small refractory period. Top: sample of network activity, rastergram of 80 randomly selected neurons. Middle: population averaged firing rate. Choice of the parameters: $N = 20000$, $C = 500$, $g = 5$, $\tau_{rp} = 0.01$ ms, $J = 0.9$ mV. Bottom: normalized variance of the population-averaged firing rate as a function of the network size (computed with 1 ms bins).

power law decay for intermediate J values.

Limits of the Gaussian approximation

A different effect is found by increasing the dominance of inhibition over excitation in the network, i.e. by increasing g, or equivalently, by decreasing f. As shown in Fig. 3 **a**, inhibition dominance can significantly deform the shape of the distribution, which displays suppressed tails for positive currents. As the inhibition dominance is increased, since $\phi(x_i)$ is positive and J_{ij} strongly negative on average, the fluctuations become increasingly skewed in the negative direction. As expected, the Gaussian approximation does not fit well the simulated data. Fig. 3 **b-c** shows that the same effect is quite general and extends to networks where excitation and inhibition are not segregated or the connectivity C is random.

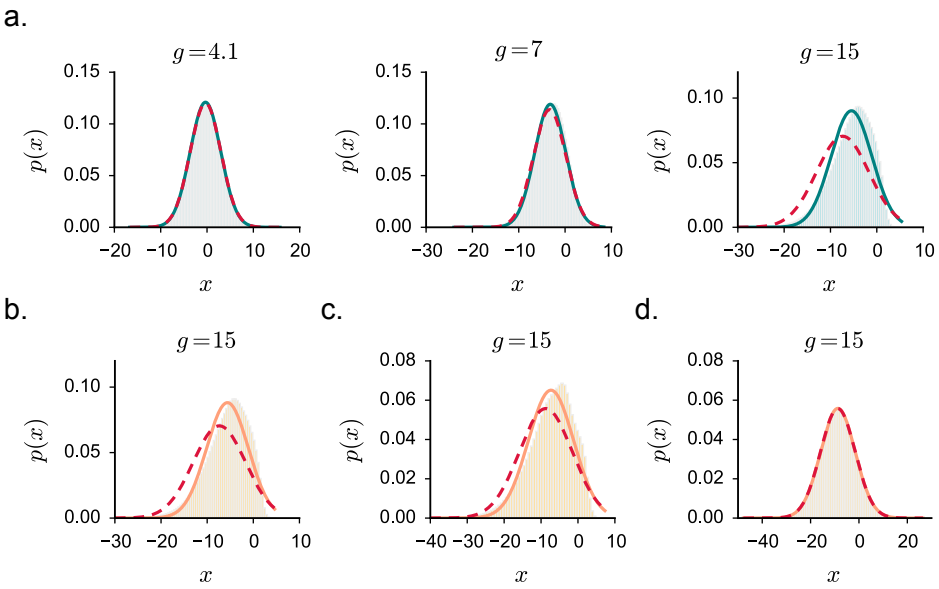


Figure 3: Comparison between dynamical mean field predictions and numerical simulations: the effects of strong inhibition. Distribution of the input current x in the population and in different time steps. The numerical distribution is obtained through averaging over 3 realizations of the synaptic matrix. Light green/orange: simulated data distribution, dark green/orange: best Gaussian fit to data, red: DMF prediction. Choice of the parameters: $C = 100$, $N = 6500$, $J = 0.2$. **a.** Dependence on the inhibition dominance g. **b.** Numerical distribution for a network with a synaptic matrix where C is fixed, as above, but excitatory and inhibitory units are shuffled. **c.** As above, with a synaptic matrix where C is random. **d.** As above, with the equivalent Gaussian matrix, whose statistics match the ones of the sparse one.

An extreme consequence of this effect is the failure of DMF in describing purely inhibitory networks in absence of external excitatory currents, where the effective coupling $\eta_i(t) = \sum_j J_{ij}\phi(x_j(t))$ is strictly non-positive at all times. In this case, DMF erroneously predicts a critical coupling J_D between a bounded and an unbounded regime, the divergence being led by the positive tails of the Gaussian bell (not shown). In contrast, in absence of any positive feedback, purely inhibitory networks cannot display a transition to run-away activity.

As a final remark, we observe that the agreement between simulated activity and mean field predictions in the case of purely inhibitory networks is in general less good than the one we found for E-I architectures (not shown).

We conclude that the Gaussian hypothesis adopted in the DMF framework is a reasonable approximation only when inhibition does not overly dominate excitation. Finally, we remark that this limitations critically depends on adopting sparse matrices where non-zero entries have fixed values. If adopting a Gaussian, fully-connected connectivity, whose mean and variance are matching the ones of the original matrix:

$$\begin{aligned}[J_{ij}] &= \frac{J}{N}(C_E - gC_I) \\ [J_{ij}^2] &= \frac{J^2}{N}(C_E + g^2 C_I)\end{aligned} \tag{1}$$

numerical simulations reveal that, whatever the degree of inhibition, positive entries are strong enough to balance the distribution, which strongly resembles again a Gaussian bell Fig. 3 **d**.

Appendix B
DMF equations in generalized E-I settings

In Chapter 2, we provide a step-by-step derivation of the Dynamical Mean Field equation in the case of the simplest excitatory-inhibitory (E-I) architecture, where the connectivity in-degree is fixed and external inputs are constant and homogeneous.

Here we relax some of these assumptions, and we provide a detailed account on how mean field equations should be modified in the case of more general E-I network models. The solutions to the new sets of equations are presented and discussed in Chapter 3.

Mean field theory in presence of noise

In order to investigate the effect of an external noisy input on the dynamical regimes, we introduced an additive, white noise term in the network dynamics equations, which read:

$$\dot{x_i}(t) = -x_i(t) + \sum_{j=1}^{N} J_{ij}\phi(x_j(t)) + \xi_i(t) \tag{2}$$

with $[\xi_i(t)] = 0$ and $[\xi_i(t)\xi_j(t+\tau)] = 2\Delta_{ext}\delta_{ij}\delta(\tau)$.

As above, we replace the forcing term $\sum_j J_{ij}\phi(x_j) + \xi_i$ by an effective noise η_i. By following the same steps as before we find:

$$\begin{aligned} &[\eta_i(t)] = J(C_E - gC_I)[\phi] \\ &[\eta_i(t)\eta_i(t+\tau)] - [\eta_i]^2 = \delta_{ij}\left[J^2(C_E + g^2C_I)\left\{C(\tau) - [\phi]^2\right\} + 2\Delta_{ext}\delta(\tau)\right] \end{aligned} \tag{3}$$

which translates into:

$$\ddot{\Delta}(\tau) = \Delta(\tau) - J^2(C_E + g^2C_I)\{C(\tau) - [\phi]^2\} + 2\Delta_{ext}\delta(\tau). \tag{4}$$

We conclude that the external noise acts on the auto-correlation function by modifying its initial condition into: $\dot{\Delta}(0^+) = -\dot{\Delta}(0^-) = -\Delta_{ext}$. In terms of the analogy with the 1D

motion, the presence of noise translates into an additive kinetic term in $\tau = 0$, which one has to take into account while writing down the energy balance:

$$V(\Delta_0, \Delta_0) + \frac{1}{2}\dot{\Delta}(0)^2 = V(0, \Delta_0) \tag{5}$$

to be solved again together with the equation for the mean μ. The potential $V(\Delta, \Delta_0)$, in contrast, remains unperturbed. The main effect of including a kinetic term at $\tau = 0$ consists in allowing a variance $\Delta_0 \neq 0$ also in the low coupling regime, where the potential has the usual shape as in Fig. 2.2 **a**.

From a mean field perspective, white noise can be studied as a proxy for the effect induced by spikes on the rate dynamics. In order to better quantify this effect, following [69], we add a spiking mechanism on the rate dynamics in Eq. 2.21. Spikes are emitted according to independent inhomogeneous Poisson processes of rate $\phi(x_j(t))$, which obeys:

$$\bar{\tau}\dot{x}(t) = -x(t) + \sum_{j=1}^{N} J_{ij}\chi_j(t) \tag{6}$$

and $\chi_j(t)$ is the spike train emitted by neuron j: $\chi_j(t) = \sum_k \delta(t - t_j^k)$.

This simple spiking mechanism can be again incorporated into a DMF description. Here, following [69], we show that the resulting equations correspond to an usual rate model with additive white noise, whose variance is given by $J^2(C_E + g^2 C_I)[\phi]/\bar{\tau}$. The mean field forcing noise is in this case $\eta_i(t) = \sum_j J_{ij}\chi_j(t)$. By separating J_{ij} into the sum of its mean and a zero-mean term, we get that the usual equation for the first order statistics holds:

$$[\eta_i] = J(C_E - gC_I)[\phi]. \tag{7}$$

In order to compute the noise auto-correlation, we separate η_i into a rate and a zero-mean spikes contribution: $\eta_i = \eta_i^r + \eta_i^{sp}$, where $\eta_i^r = \sum_j J_{ij}\phi(x_j)$ and $\eta_i^{sp} = \sum_j J_{ij}\{\chi_j - \phi(x_j)\}$. The auto-correlation of the rate component returns the usual contribution:

$$[(\eta_i^r(t) - [\eta_i^r])(\eta_j^r(t+\tau) - [\eta_j^r])] = \delta_{ij}J^2(C_E + g^2C_I)\{C(\tau) - [\phi]^2\} \tag{8}$$

while the auto-correlation of the spikes term generates the instantaneous variability induced by the Poisson process:

$$[(\eta_i^{sp}(t) - [\eta_i^{sp}])(\eta_j^{sp}(t+\tau) - [\eta_j^{sp}])] = \delta_{ij}J^2(C_E + g^2C_I)[\phi]\delta(\tau). \tag{9}$$

By summing the two contributions together, and rescaling time appropriately, we obtain the evolution equation for $\Delta(\tau)$ equivalent to Eq. 4 with a self-consistent white noise term:

$$\ddot{\Delta}(\tau) = \Delta(\tau) - J^2(C_E + g^2C_I)\{C(\tau) - [\phi]^2 + \frac{[\phi]}{\bar{\tau}}\delta(\tau)\}. \tag{10}$$

Mean field theory with stochastic in-degree

We derive here the dynamical mean field equations for networks in which the total number of inputs C varies randomly between different units in the network. We focus on a connectivity

matrix with one excitatory and one inhibitory column. In the excitatory column, each element J_{ij} is drawn from the following discrete distribution:

$$J_{ij} = \begin{cases} J & p = C_E/N_E = C/N \\ 0 & \text{otherwise.} \end{cases}$$

Up to the the order $O(1/N)$, the statistics of the entries J_{ij} are are:

$$[J_{ij}] = \frac{J}{N} C, \tag{11}$$

$$[J_{ij}^2] = \frac{J^2}{N} C. \tag{12}$$

The inhibitory column is defined in a similar way, if substituting J with $-gJ$.

We proceed in the same order as in the previous sections. We define the effective stochastic coupling, given by $\eta_i(t) = \sum_j J_{ij}\phi(x_j(t))$. We compute the equations for the mean and the correlation of the Gaussian noise η_i in the thermodynamic limit.

We will find that the variance associated to the single neuron activity will consist of a temporal component, coinciding with the amplitude squared of chaotic fluctuations, and of a quenched term, which appears when sampling different realizations of the random connectivity matrix.

For a given realization and a given unit i, the temporal auto-correlation coincides with: $[\eta_i(t)\eta_i(t+\tau)]_{t,ic} - [\eta_i]^2_{t,ic}$ by averaging over time and over different initial conditions. In a second step, averaging over all the units in the population, or equivalently, over the realizations of the matrix J_{ij}, returns the average size of deviations from single unit mean within one single trial $[[\eta_i(t)\eta_i(t+\tau)]_{t,ic} - [\eta_i]^2_{t,ic}]_J = [\eta_i(t)\eta_i(t+\tau)] - [[\eta_i]^2_{t,ic}]_J$. Remember that, in our notation, [] indicates an average over time, initial conditions, and matrix realizations. One can compute self-consistently this quantity and check that it coincides with the expression for the total second order moment we found in the previous paragraph for the fixed in-degree case.

In order to close the expression for the DMF equations, we will need to express all the averages of ϕ in terms of the total variance Δ_0, which includes quenched variability. For this reason, we compute the average deviations from $[\eta_i(t)\eta_i(t+\tau)]$ with respect to the population mean $[\eta_i]$. As a result, the second moment $[\eta_i(t)\eta_j(t+\tau)] - [\eta_i(t)]^2$ will now include the static trial-to-trial variability.

For the mean, we get:

$$\begin{aligned} [\eta_i(t)] &= \left[\sum_{j_E=1}^{N_E} J_{ij_E}\phi_{j_E}(t)\right] + \left[\sum_{j_I=1}^{N_I} J_{ij_I}\phi_{j_I}(t)\right] = \left(N_E\left[J_{ij_E}\right] + N_I\left[J_{ij_I}\right]\right)[\phi] \\ &= J(C_E - gC_I)[\phi]. \end{aligned} \tag{13}$$

Applying the same steps as before, we compute the second order statistics:

$$\begin{aligned}[\eta_i(t)\eta_i(t+\tau)] &= \left[\sum_{k=1}^{N} J_{ik}\phi_k(t)\sum_{l=1}^{N} J_{il}\phi_l(t+\tau)\right] \\ &= \left[\sum_{k_E=1}^{N_E} J_{ik_E}\phi_{k_E}(t)\sum_{l_E=1}^{N_E} J_{il_E}\phi_{l_E}(t+\tau)\right] + \left[\sum_{k_I=1}^{N_I} J_{ik_I}\phi_{k_I}(t)\sum_{l_I=1}^{N_I} J_{il_I}\phi_{l_I}(t+\tau)\right] \\ &+ 2\left[\sum_{k_E=1}^{N_E} J_{ik_E}\phi_{k_E}(t)\sum_{l_I=1}^{N_I} J_{il_I}\phi_{l_I}(t+\tau)\right].\end{aligned} \tag{14}$$

Again, we consider separate contributions from diagonal ($k = l$) and off-diagonal ($k \neq l$) terms. This results in:

$$\begin{aligned}&[\eta_i(t)\eta_i(t+\tau)] = C_E J^2[\phi_i(t)\phi_i(t+\tau)] + C_E^2(1-1/N_E)J^2[\phi]^2 \\ &- 2C_E C_I g J^2[\phi]^2 + C_I g^2 J^2[\phi_i(t)\phi_i(t+\tau)] + C_I^2(1-1/N_I)g^2J^2[\phi]^2.\end{aligned} \tag{15}$$

As we can see, diagonal terms behave, on average, like in the fixed in-degree case. To estimate the off-diagonal contributions, we observe that for every k_E index, the expected number of other non-zero incoming connections is $C_E(1-1/N_E)$. As a consequence, the $k_E \neq l_E$ sum contains on average C_E^2 terms of value $J^2\,[\phi]^2$ in the limit $N \to \infty$. Note that in the fixed in-degree case, the same sum contained exactly $C_E(C_E-1)$ terms. That resulted in a smaller value for the second order statistics, which does not include the contribution from stochasticity in the number of incoming connections. Similar arguments hold for the inhibitory units.

To conclude, in the large network limit, we found:

$$[\eta_i(t)\eta_i(t+\tau)] = J^2(C_E + g^2C_I)\langle\phi_i(t)\phi_i(t+\tau)\rangle + J^2(C_E - gC_I)^2\,[\phi]^2 \tag{16}$$

such that the final result reads:

$$[\eta_i(t)\eta_i(t+\tau)] - [\eta_i(t)]^2 = J^2(C_E + g^2C_I)C(\tau). \tag{17}$$

As before, one can then check that the cross-correlation between different units vanishes.

The noise distribution determines the following self-consistent potential:

$$V(\Delta,\Delta_0) = -\frac{\Delta^2}{2} + J^2(C_E+g^2C_I)\int \mathcal{D}z\left[\int \mathcal{D}x\Phi(\mu+\sqrt{\Delta_0-|\Delta|}x+\sqrt{|\Delta|}z)\right]^2. \tag{18}$$

In contrast with the potential of Eq. 2.35, which was found for networks with fixed in-degree, here we observe the lack of the term $-\Delta[\phi]^2$. As a consequence, the new potential is flat around a non-zero $\Delta = \Delta_\infty$ value, which represents the asymptotic population disorder.

As usually, we derive the DMF solution in the weak and in the strong coupling regime thanks to the analogy with the one-dimensional equation of motion. When $J < J_C$, the potential has the shape of a concave parabola, the vertex of which is shifted to $\Delta_\infty \neq 0$. The only acceptable physical solution is here $\Delta(\tau) = \Delta_0 = \Delta_\infty$. In order to determine its value, we use the condition emerging from setting $\ddot{\Delta} = 0$:

$$\Delta_0 = J^2(C_E+g^2C_I)\int \mathcal{D}z\phi^2(\mu+\sqrt{\Delta_0}z) \tag{19}$$

to be solved together with the equation for the mean:

$$\mu = J(C_E - gC_I) \int \mathcal{D}z \phi(\mu + \sqrt{\Delta_0} z). \tag{20}$$

When $J > J_C$, the auto-correlation acquires a temporal structure. The stable solution is monotonically decreasing from Δ_0 to a value Δ_∞, and we need to self-consistently determine μ, Δ_∞ and Δ_0 through three coupled equations. Apart from the usual one for μ, a second equation is given by the energy conservation law:

$$V(\Delta_0, \Delta_0) = V(\Delta_\infty, \Delta_0) \tag{21}$$

which reads:

$$\frac{\Delta_0^2 - \Delta_\infty^2}{2} = J^2(C_E + g^2 C_I) \left\{ \int \mathcal{D}z \Phi^2(\mu + \sqrt{\Delta_0} z) - \int \mathcal{D}z \left[\int \mathcal{D}x \Phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2 \right\}. \tag{22}$$

The third equation emerges from setting $\ddot{\Delta} = 0$ at Δ_∞, which gives:

$$\Delta_\infty = J^2(C_E + g^2 C_I) \int \mathcal{D}z \left[\int \mathcal{D}x \phi(\mu + \sqrt{\Delta_0 - \Delta_\infty} x + \sqrt{\Delta_\infty} z) \right]^2. \tag{23}$$

Mean field theory in general E-I networks

We discuss here the more general case of a block connectivity matrix, corresponding to one excitatory and one inhibitory population receiving statistically different inputs. The synaptic matrix is now given by:

$$\mathrm{J} = J \left(\begin{array}{c|c} \mathrm{J_{EE}} & \mathrm{J_{EI}} \\ \hline \mathrm{J_{IE}} & \mathrm{J_{II}} \end{array} \right). \tag{24}$$

Each row of J contains exactly C_E non-zero excitatory entries in the blocks of the excitatory column, and exactly C_I inhibitory entries in the inhibitory blocks. Non-zero elements are equal to j_E in $\mathrm{J_{EE}}$, to $-g_E j_E$ in $\mathrm{J_{EI}}$, to j_I in $\mathrm{J_{IE}}$, and to $-g_I j_I$ in $\mathrm{J_{II}}$.

The network admits a fixed point (x_0^E, x_0^I) which is homogeneous within the two different populations:

$$\begin{pmatrix} x_0^E \\ x_0^I \end{pmatrix} = J \begin{pmatrix} j_E(C_E \phi(x_0^E) - g_E C_I \phi(x_0^I)) \\ j_I(C_E \phi(x_0^E) - g_I C_I \phi(x_0^I)) \end{pmatrix}. \tag{25}$$

With linear stability analysis, we obtain that the fixed point stability is determined by the eigenvalues of matrix:

$$\mathrm{S} = J \left(\begin{array}{c|c} \phi_E' \mathrm{J_{EE}} & \phi_I' \mathrm{J_{EI}} \\ \hline \phi_E' \mathrm{J_{IE}} & \phi_I' \mathrm{J_{II}} \end{array} \right) \tag{26}$$

where we used the short-handed notation $\phi_k' = \phi'(x_0^k)$.

The eigenspectrum of S consists of a densely distributed component, represented by a circle in the complex plane, and a discrete component, consisting of two outlier eigenvalues. The

radius of the complex circle is determined by the 2×2 matrix containing the variance of the entries distributions in the four blocks, multiplied by N [6, 7, 5]:

$$\Sigma = J^2 \begin{pmatrix} {\phi'_E}^2 C_E j_E^2 & {\phi'_I}^2 C_I g_E^2 j_E^2 \\ {\phi'_E}^2 C_E j_I^2 & {\phi'_I}^2 C_I g_I^2 j_I^2 \end{pmatrix}. \tag{27}$$

More precisely, the radius of the circle is given by the square root of its larger eigenvalues:

$$\begin{aligned} r = \Bigg[\frac{1}{2} J^2 \Big\{ & C_E {\phi'_E}^2 j_E^2 + C_I {\phi'_I}^2 g_I^2 j_I^2 \\ & + \sqrt{(C_E {\phi'_E}^2 j_E^2 + C_I {\phi'_I}^2 g_I^2 j_I^2)^2 - 4 C_E C_I {\phi'_E}^2 {\phi'_I}^2 j_E^2 j_I^2 (-g_E^2 + g_I^2)} \Big\} \Bigg]^{\frac{1}{2}} \end{aligned} \tag{28}$$

where the derivative terms ϕ'_k contain an additional dependency on J.

In order to determine the two outlier eigenvalues, we construct the 2×2 matrix containing the mean of S in each of the four blocks, multiplied by N:

$$\mathrm{M} = J \begin{pmatrix} \phi'_E C_E j_E & -\phi'_I C_I g_E j_E \\ \phi'_E C_E j_I & -\phi'_I C_I g_I j_I \end{pmatrix}. \tag{29}$$

The outliers correspond to the two eigenvalues of M, and are given by:

$$\xi_\pm = \frac{1}{2} J \left\{ \phi'_E C_E j_E - \phi'_I C_I g_I j_I \pm \sqrt{(\phi'_E C_E j_E - \phi'_I C_I g_I j_I)^2 + 4 \phi'_E \phi'_I C_E C_I j_E j_I (-g_E + g_I)} \right\}. \tag{30}$$

Notice that, if g_E is sufficiently larger than g_I, the outlier eigenvalues can be complex conjugates.

We focus on the case where, by increasing the global coupling J, the instability to chaos is the first bifurcation to take place. As in the simpler case when excitatory and inhibitory populations are identical, we need the real part of the outliers to be negative or positive but smaller than the radius r of the densely distributed component of the eigenspectrum. This requirement can be accomplished by imposing relative inhibitory strengths g_E and g_I strong enough to overcome excitation in the network. For a connectivity matrix which satisfies the conditions above, an instability to a fluctuating regime occurs when the radius r crosses unity.

We can use again DMF to analyze the network activity below the instability. To start with, dealing with continuous-time dynamics, one can easily generalize the mean field equations we recovered for the simpler two-column connectivity. In the new configuration, the aim of mean field theory is to determine two values of the mean activity and two values for the variance, one for each population.

By following the same steps as before, we define $\eta_i^E = \sum_{j=1}^N J_{ij} \phi(x_j(t))$ for each i belonging to the E population, and $\eta_i^E = \sum_{j=1}^N J_{ij} \phi(x_j(t))$ for each i belonging to I. Those two variables represent the effective stochastic inputs to excitatory or inhibitory units which replace the deterministic network interactions. Under the same hypothesis as before, we compute the statistics of the η_i^E and η_i^I distributions. For the mean, we find:

$$\begin{pmatrix} [\eta_i^E] \\ [\eta_i^I] \end{pmatrix} = J \begin{pmatrix} C_E j_E & -C_I g_E j_E \\ C_E j_I & -C_I g_I j_I \end{pmatrix} \begin{pmatrix} [\phi^E] \\ [\phi^I] \end{pmatrix}. \tag{31}$$

For the second order statistics, we have:

$$\begin{pmatrix} \left[(\eta_i^E(t)-[\eta_i^E])(\eta_j^E(t+\tau)-[\eta_j^E])\right] \\ \left[(\eta_i^I(t)-[\eta_i^I])(\eta_j^I(t+\tau)-[\eta_j^I])\right] \end{pmatrix} = J^2 \begin{pmatrix} C_E j_E^2 & C_I g_E^2 j_E^2 \\ C_E j_I^2 & C_I g_I^2 j_I^2 \end{pmatrix} \begin{pmatrix} C^E(\tau) - \left[\phi^E\right]^2 \\ C^I(\tau) - \left[\phi^I\right]^2 \end{pmatrix}. \tag{32}$$

By using those results, we obtain two equations for the mean values of the input currents:

$$\begin{pmatrix} \mu^E \\ \mu^I \end{pmatrix} = J \begin{pmatrix} C_E j_E & -C_I g_E j_E \\ C_E j_I & -C_I g_I j_I \end{pmatrix} \begin{pmatrix} \left[\phi^E\right] \\ \left[\phi^I\right] \end{pmatrix} \tag{33}$$

and two differential equations for the auto-correlation functions, which can be summarized as:

$$\begin{pmatrix} \ddot{\Delta}^E(\tau) \\ \ddot{\Delta}^I(\tau) \end{pmatrix} = \begin{pmatrix} \Delta^E(\tau) \\ \Delta^I(\tau) \end{pmatrix} - J^2 \begin{pmatrix} C_E j_E^2 & C_I g_E^2 j_E^2 \\ C_E j_I^2 & C_I g_I^2 j_I^2 \end{pmatrix} \begin{pmatrix} C^E(\tau) - \left[\phi^E\right]^2 \\ C^I(\tau) - \left[\phi^I\right]^2 \end{pmatrix}. \tag{34}$$

All the mean values are defined and computed as before, the population averages to be taken only over the E or the I population.

The main difficulty in solving Eqs. 33 and 34 comes from the absence of an analogy with an equation of motion for a classical particle in a potential. Unfortunately, indeed, isolating the self-consistent solution in absence of an analogous suitable potential $V(\Delta^E(\tau), \Delta^I(\tau))$ appears to be computationally very costly.

However, if we restrict ourselves to discrete-time rate dynamics:

$$x_i(t+1) = \sum_{j=1}^{N} J_{ij}\phi(x_j(t)). \tag{35}$$

DMF equations can still easily be solved. With discrete-time evolution, the mean field dynamics reads:

$$x_i(t+1) = \eta_i(t) \tag{36}$$

which identifies directly the input current variable x_i with the stochastic process η_i. In contrast to the continuous case, where self-consistent noise is filtered by a Langevin process, the resulting dynamics is extremely fast. As a consequence, the statistics of η_i directly translates into the statistics of x. We are left with four variables, to be determined according to four equations, which can be synthesized in the following way:

$$\begin{pmatrix} \mu_E \\ \mu_I \end{pmatrix} = J \begin{pmatrix} C_E j_E & -C_I g_E j_E \\ C_E j_I & -C_I g_I j_I \end{pmatrix} \begin{pmatrix} \left[\phi^E\right] \\ \left[\phi^I\right] \end{pmatrix} \tag{37}$$

$$\begin{pmatrix} \Delta_0^E \\ \Delta_0^I \end{pmatrix} = J^2 \begin{pmatrix} C_E j_E^2 & C_I g_E^2 j_E^2 \\ C_E j_I^2 & C_I g_I^2 j_I^2 \end{pmatrix} \begin{pmatrix} \left[\phi^{E2}\right] - \left[\phi^E\right]^2 \\ \left[\phi^{I2}\right] - \left[\phi^I\right]^2 \end{pmatrix}. \tag{38}$$

As usual, firing rate statistics are computed as averages with respect to a Gaussian distribution with mean μ_E (μ_I) and variance Δ_0^E (Δ_0^I).

When adopting discrete-time dynamics, a second condition has to be imposed on the connectivity matrix. To prevent phase-doubling bifurcations specific to discrete-time dynamics, we need the real part of the outliers to be strictly smaller than r in modulus. An isolated outlier

on the negative real axis, indeed, would lose stability and induce fast oscillations in the activity before the transition to chaos takes place. The latter condition is satisfied in a regime where inhibition is only weakly dominating, coinciding with the phase region where the approximation provided by DMF is very good (see in Appendix A).

Appendix C
Unit rank structures in networks with positive activation functions

In Chapters 7, 8 and 9, we performed our analysis of partially structured networks by adopting a completely symmetric network model, whose input-free solutions are invariant under the sign transformation $x_i(t) \rightarrow -x_i(t)$. As it was shown in Section 7.3.8, symmetry can be broken by including additional external input currents. Another possibility, which brings the network closer to biologically inspired circuit models, is to adopt a non symmetric, positively defined activation function $\phi(x)$.

Here, we specifically investigate the effect of changing the transfer function to: $\phi(x) = 1 + \tanh(c(x-\gamma))$. Adding a shift γ is equivalent to include an external and constant negative input. The parameter c, instead, rescales the slope of $\phi(x)$ at the inflection point.

For simplicity, we fix $\gamma = 1$ and $c = 1.5$. We furthermore restrict the analysis to the case of unit rank structures whose right- and left-structure vectors solely overlap on the unitary direction ($\rho = 0$).

In absence of any disorder ($g = \Sigma_m = 0$), the fixed point equation reads: $x^0 = M_m M_n \phi(x^0)$. The unstable fixed point thus coincides with $x^0 = 1$, while the two stable ones are built on the high and low firing rate branches of $\phi(x)$. In contrast to the symmetric case we studied so far, a modulation in $M_m M_n$ changes both the maximal slope slope and the central intersection of $\phi(x)$ with the bisector. As a consequence, when $M_m M_n \gtrsim 1$, the central fixed point moves to small firing rate values. For $M_m M_n \gg 1$ it finally merge with the low firing state, so that only one high-firing rate fixed point exists. Similarly, when $M_m M_n \lesssim 1$, the unstable fixed point moves towards the high firing one, before annihilating with it and disappearing. In this regime, one unique low-firing rate state exists.

Dynamical Mean Field solutions

When the network solutions are not homogeneous, the behaviour of the static and chaotic solutions needs to be studied with the usual mean field tools. The Dynamical Mean Field (DMF) sets of equations 7.72 and 7.66 were derived for an arbitrary activation function, so

they can directly be adopted in the present scenario. We start from graphically analysing the stationary solutions in Eq. 7.72, and we plot the two nullclines of the system for different values of the architecture parameters.

Fig. 4 **a** (left) displays the μ nullclines for different $M_m M_n$ values. The result is in agreement with the simple picture we derived in the case of homogeneous fixed points. At $M_m M_n = 1$, the unstable branch coincides with $\mu = 1$, and the stable ones are symmetric. Around $M_m M_n = 1$, the perfect pitchfork is broken in one or the other direction, generating a first stable continuous branch and a second one, where one unstable and one stable solution merge at low or high firing rate. For extremely low (high) $M_m M_n$ values, finally, there's just one nullcline at low (high) μ values.

The Δ_0 nullcline (Fig. 4 **a**, right) displays a more complex behaviour with respect to the symmetric $\phi(x) = \tanh(x)$ case. When g is sufficiently large, indeed, it can become a non-monotonic function of the mean input μ, transforming into a *S*-shaped nullcline. As it will be shown in detail, this more complex shape is able to induce bistable activity even when the μ nullcline is reduced to a single continuous branch. This situation is reminiscent of the *fluctuations driven* bistable regime in [103].

We find that the system admits two classes of stable solutions (Fig. 4 **b-c**). The first one, plotted in Fig. 4 **b**, takes large mean and variance values. It suddenly disappears on the leftmost grey boundary of the plot, in a parameter region which co-exist with the second solution. The latter solution, plotted in Fig. 4 **c**, takes typically small values of μ and Δ_0, and disappears on the rightmost boundary with a first-order transition as well.

In order to dissect more systematically the nature of those solutions, and the kind of bifurcations taking place on the stability boundaries, we imagine to fix the structure strength (dashed lines in Fig. 4 **b**), and to gradually increase the random strength g.

First, in Fig. 5, we fix the structure strength to high values: $M_m M_n = 1.2$. The bifurcation pattern occurring in this case resembles what we observed in the original case with $\phi(x) = \tanh(x)$. At low values of g, two stable fixed points are built, respectively, on the high and on the low branches of the μ nullcline. For that reason, we call this state LH (cfr Fig. 8). When the random strength is too strong, the low firing rate fixed point annihilates, and only one high firing solution survives (H state). Such solution finally smoothly transforms into a chaotic one. Both instabilities are correctly predicted by our estimation of the compact and the discrete components of the stability eigenspectrum S_{ij}.

When $M_m M_n$ is exactly equal to unity (Fig. 6), the nullcline for μ is a perfectly symmetric pitchfork. At small g values, similarly to the previous case, network activity is bistable and admits one L and one H stationary state. As g increases, the Δ_0 intersect the high firing rate branch at smaller and smaller values of μ. Finally, the H state is lost, and the second stable fixed point is realized on the intermediate branch at $\mu = 1$. This bistable state is thus formally a LI state. Finally, at large g values, the two intersections on the low rate branch collapse together and disappear. Bistability is lost and only one intermediate (I) state exist.

The intermediate branch of the μ nullcline exists only when $M_m M_n$ is exactly equal to unity. For this reason, I states are represented in phase diagram regions with null measure (dashed line in Fig. 8). However, I states separate the phase diagram in two macro areas: below the dashed line, every stationary and chaotic solution is build on the same low firing rate branch of the μ nullcline, and is thus formally a L state.

When only L states are present, bifurcations are discontinuous and *S*-shaped. They can be observed for slightly smaller values of the structure strength: in Fig. 7, we fix $M_m M_n = 0.98$. In this case, while a classical LH state exists at small g values, the bistable state at large random

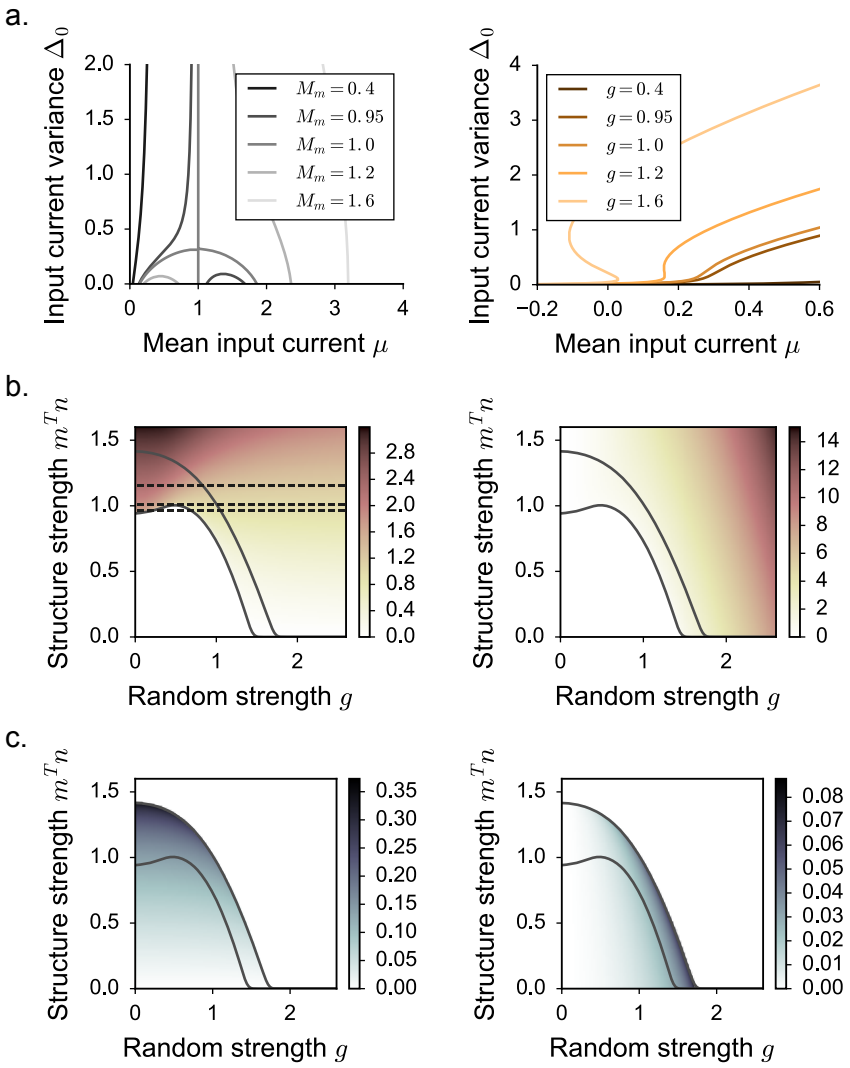


FIGURE 4: Dynamical Mean Field stationary solutions in network models with non-symmetric, positively defined activation functions. **a.** Nullclines for the system of equations in 7.72. Left: μ nullclines for different values of the structure strength parameter $M_m M_n$. Right: Δ_0 nullclines for different values of the random strength g. **b-c.** Stationary stable solutions plotted as color maps on the parameter space defined by the random and the structure strengths. The two main classes of continuous solutions are displayed, respectively, in panels **b** and **c**. Left: μ, right: Δ_0. Dark grey continuous lines: boundaries of the parameter region where both stable solutions exist. Horizontal dashed lines: values of the structure strength used for the bifurcation analysis in Fig. 5, 6 and 7.

strengths involves two stable solutions which originate both a low firing rates (LL state). The two states strongly differ in the value of their variance. When g is sufficiently large, one unique low firing rate, high variance state survives.

All the different activity states are finally sketched in the phase diagram of Fig. 8. The exact shape of the phase diagram depends on the value of the parameters c and γ. Note that

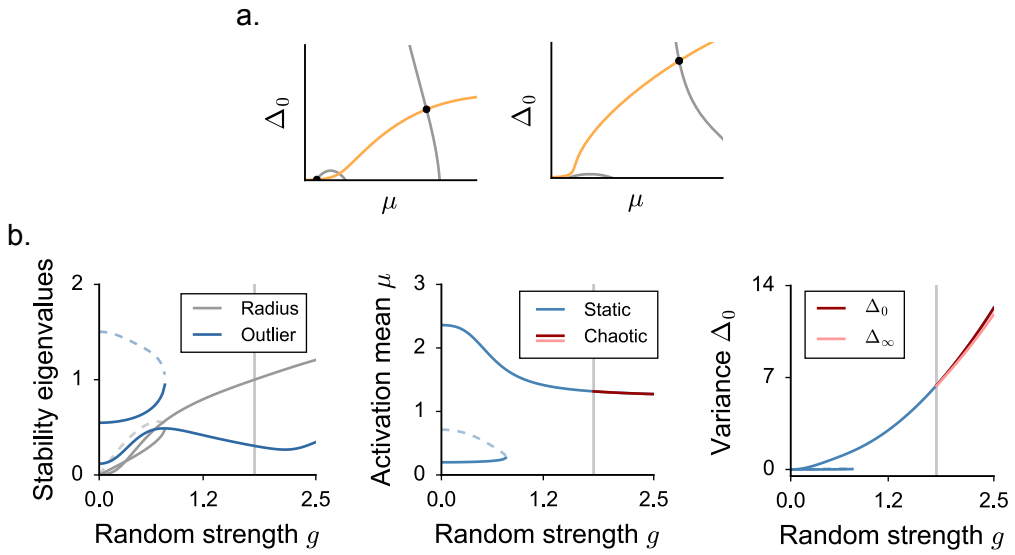


Figure 5: Dynamical Mean Field stationary solutions in network models with $\phi(x) = 1 + \tanh(c(x-\gamma))$. The structure strength is fixed to 1.2 (top dashed line in Fig. 4 **b**). **a.** Graphical analysis of the system of equations in 7.72 for $g = 0.4$ and $g = 1$. Details as in Fig. 7.4. **b.** Bifurcation diagrams for increasing values of the random strength g.

the LL bistability region can disappear from the phase diagram when the two parameters c and γ take too small or too large values.

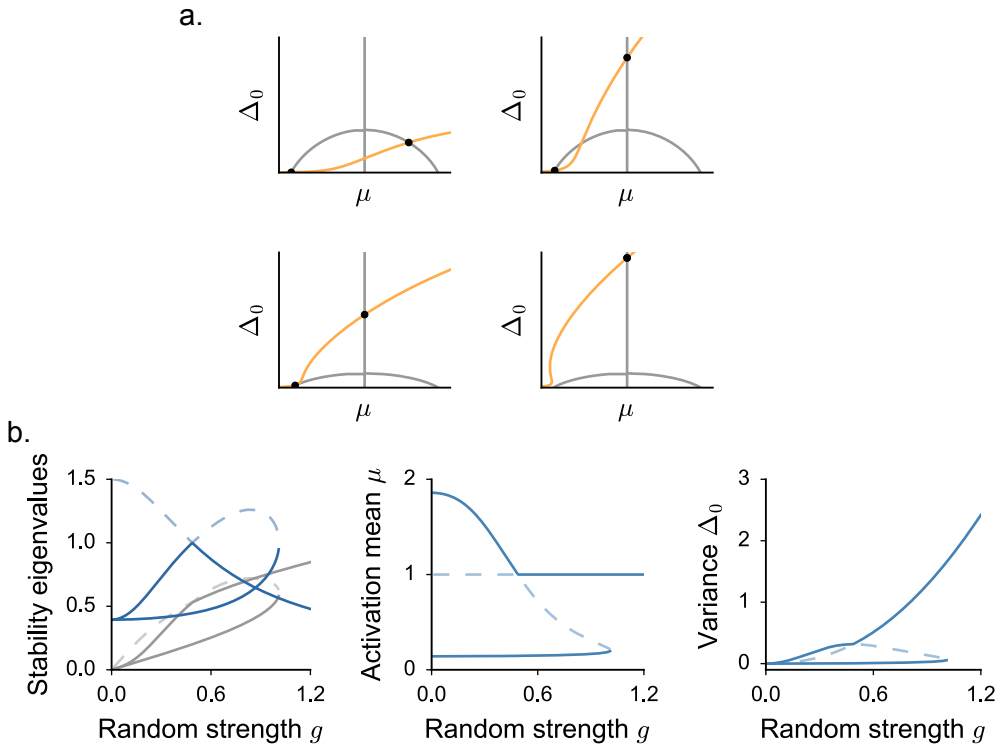


FIGURE 6: Dynamical Mean Field stationary solutions in network models with $\phi(x) = 1 + \tanh(c(x - \gamma))$. The structure strength is fixed to 1 (center dashed line in Fig. 4 **b**). **a.** Graphical analysis of the system of equations in 7.72 for $g = 0.3, 0.75, 1.0$ and 1.3. Details as in Fig. 7.4. **b.** Bifurcation diagrams for increasing values of the random strength g.

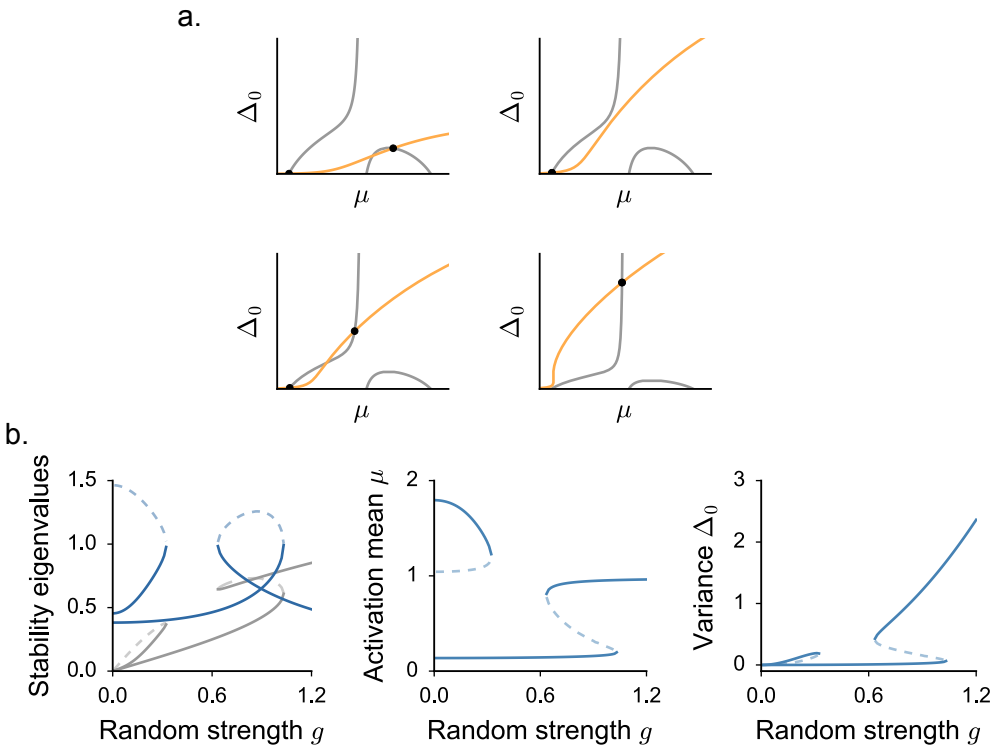


Figure 7: Dynamical Mean Field stationary solutions in network models with $\phi(x) = 1 + \tanh(c(x - \gamma))$. The structure strength is fixed to 0.98 (bottom dashed line in Fig. 4 **b**). **a.** Graphical analysis of the system of equations in 7.72 for $g = 0.3$, 0.6, 0.7 and 1.2. Details as in Fig. 7.4. **b.** Bifurcation diagrams for increasing values of the random strength g.

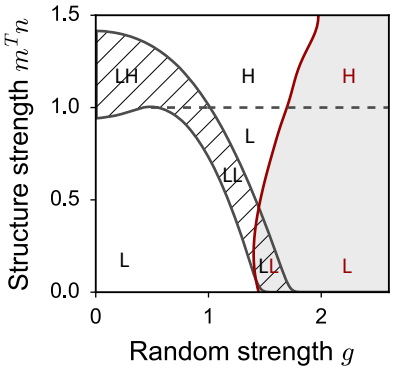


Figure 8: Phase diagram of activity in network models with $\phi(x) = 1 + \tanh(c(x - \gamma))$. Activity states are classified as L (H) if they are built on the low (high) firing rate branch of the μ nullcline. Hatched area: phase space region where activity is bistable. Dashed line: phase space region where the H solution transforms into L, smoothly passing through an intermediate I solution. Red line: instability to chaos of the high variance solution. Shaded area: the high variance solution is chaotic.

Appendix D
Two time scales of fluctuations in networks with unit rank structure

In Chapter 7, we studied the dynamics of large networks whose connectivity matrix includes a unit rank structured term. We found that, when the structure strength is large, the DMF theory predicts two stable states, which can be stationary or chaotic.

Both in stationary and chaotic bistable solutions, the population-averaged statistics of the activation variable x_i are stationary. When activity is chaotic, indeed, irregular temporal fluctuations are decorrelated from one unit to the other, so that the central limit theorem applies at every time step, and the network statistics are constant in time.

In finite size networks, however, the network statistics are not stationary. Their dynamics display instead two different time scales. The instantaneous population-averaged activity undergoes small fluctuations of amplitude $\mathcal{O}(1/\sqrt{N})$, whose time scale reflects the relaxation decay of chaotic activity. When two chaotic attractors exist, furthermore, the mean activation displays also sharp transitions from positive to negative values and viceversa (Fig. 9 **a**). Those sudden jumps correspond to global transitions from one stable attractor to the other, which are made possible by the self-sustained temporal fluctuations.

We first consider transition events as point processes, and we measure the average transition rate. We arbitrarily define a transition point as the time step at which the population-averaged activation crosses zero. The transition rate thus depends on the amplitude of finite-size fluctuations measured with respect to the average phase space distance between the two attractors. As a consequence, we expect the transition rate to depend on the architecture parameters and on the network size, but also to vary strongly from one realization of the connectivity matrix to the other.

Consistently with our DMF description, we find that transitions between attractors become rarer and rarer as the network size N is increased (Fig. 9 **b**).

We then measure the Fano factor to evaluate the variability in the count of transition events. Fig. 9 **c** reveals that, quite robustly with respect to the system size N, the average Fano factor noisily oscillates around 1.

In a second step, we numerically analyse the two different time scales of network activity

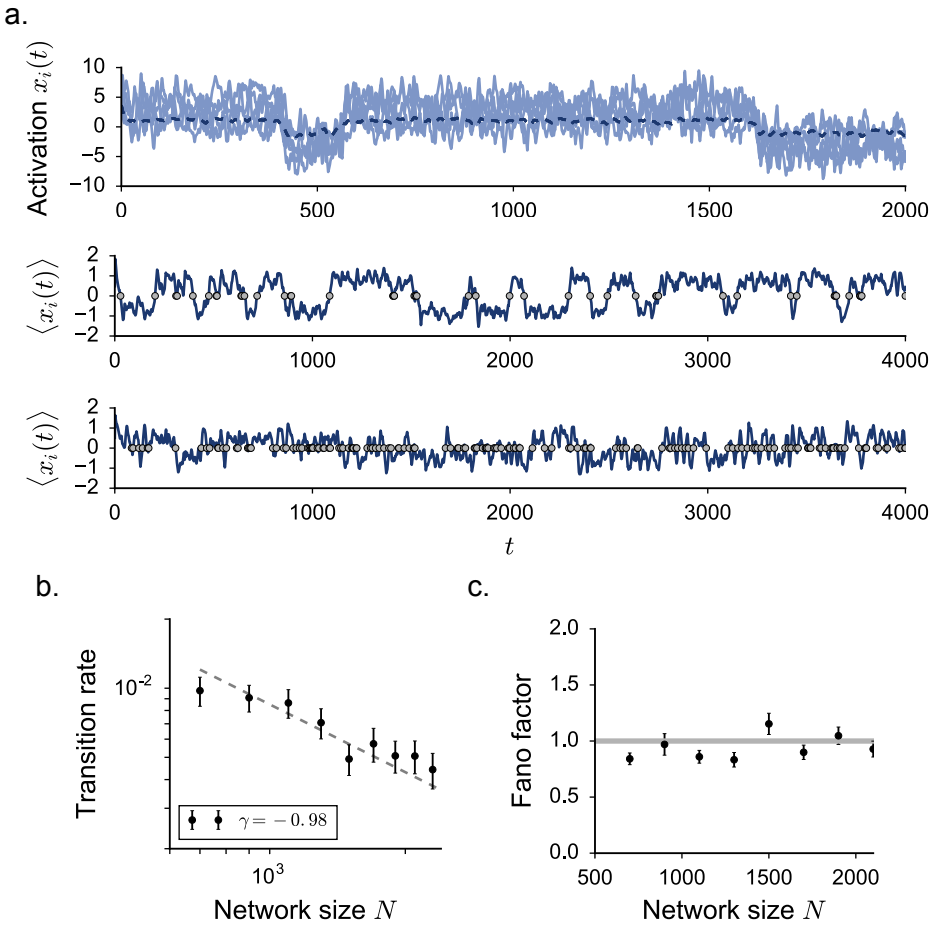


Figure 9: Chaotic activity in finite-size networks reveals transitions between the two bistable symmetric attractors. **a.** Samples of activity displaying attractors jumps. Top: activation variable for five randomly selected units (light blue). Transitions occur at the network level: at the transition point, every unit jumps from one attractor to the other. Dashed blue line: time-dependent population average. Middle and bottom: time-dependent population average in two different trials. The mean activation displays small finite-size fluctuations together with larger excursions associated with the transitions from one attractor to the other (grey points). **b.** The transition rate decays to zero as the network size N is increased. Dashed lines: power-law best fit. **c.** Fano factor of the transition point process for different values of the network size N. For every realization of the network, the jumps count is measured in different windows of the total integration time $T = 15.000$. The Fano factor is measured for every realization and then averaged over $N_{tr} = 30$ different networks. Choice of the parameters: $\rho = 0$, $g = 3.$, $M_m M_n = 3.6$, $\Sigma_m = 0$.

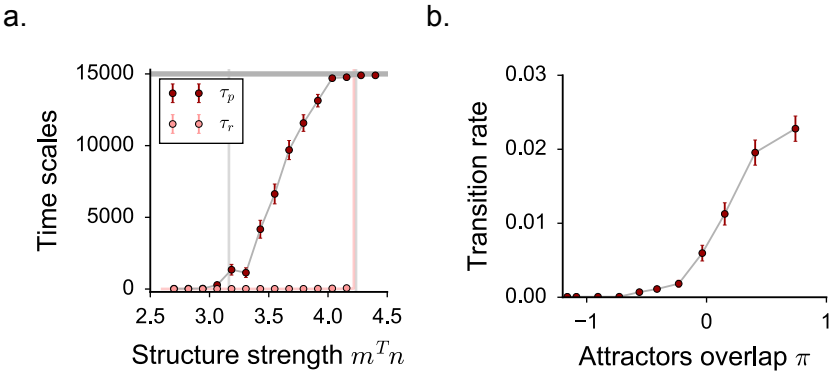


FIGURE 10: Dynamics of finite-size networks in the chaotic state are effectively characterized by two distinct time scales. **a.** Both time scales depend on the network architecture parameters. Here, we fix the random strength $g = 3$ and we increase the structure strength. Below the leftmost vertical grey line, DMF predicts an homogeneous chaotic regime; above the rightmost vertical line, two heterogeneous stationary states. Transitions between the two attractors are expected instead in the intermediate region. The average persistence time τ_p, measured as the average time interval between two transitions, grows with the structure strength, and reaches a plateau corresponding to the total simulation time (horizontal grey line) close to the transition to stationary states. The second time scale τ_r is given by the relaxation time scale of chaotic fluctuations. Pink line: DMF prediction, measured as the full width half maximum of the auto-correlation function $\Delta(\tau)$. Pink dots: a rough estimate of τ_r from finite size networks is obtained by rectifying the population average signal and we computing the full width half maximum of its auto-correlation function. **b.** The transition rate grows monotonically with the average overlap, measured from within the DMF framework. Choice of the parameters: $\rho = 0$, $g = 3.$, $\Sigma_m = 0$, $N = 1300$.

as the strength of the structured component of the connectivity is increased (Fig. 10 **a**).

The first dynamical scale is given by the relaxation time constant (τ_r), which coincides with the time course of chaotic fluctuations. Its value can be derived within the DMF framework by computing the time decay of the full auto-correlation function $\Delta(\tau)$. The second time scale is the persistence time constant (τ_p), coinciding with the average time interval separating two attractors transitions.

When the structure is weak (left region of Fig. 10 **a**), the network is in the classical homogeneous chaotic state [127]. The persistence time scale coincides here with the relaxation time constant of chaotic fluctuations. When the structured and the random components have comparable strengths, instead, two heterogeneous chaotic phases co-exist (middle region of Fig. 10 **a**). In this regime, the average persistence time increases monotonically with the structure strength. The relaxation time undergoes a very slow increase before sharply diverging at the boundary with stationary states, but the increase takes place on a much smaller scale. Finally, if the structure is too strong (right of Fig. 10 **a**), the two bistable states become stationary. In this region, τ_r is formally infinite, while τ_p coincides with the total duration of our simulations.

The increase of the persistence length with the structure strength can be linked to the increase in the phase space distance between the two attractors, centered respectively in μ and

$-\mu$. Single units trajectories are centered in $\mu + \sqrt{\Delta_\infty}$ and $-\mu - \sqrt{\Delta_\infty}$, and span in time a phase space region of typical radius $\sqrt{\Delta_0 - \Delta_\infty}$. If this radius is large enough with respect to μ, the two attractors significantly overlap and the network activity is likely to explore both trajectories during the same trial because of self-sustained disordered fluctuations.

We propose a measure for the population-averaged overlap π, and we check that it correlates with the transition frequency that we observe in finite size networks. For every unit, the typical overlap between its positive and its negative trajectories is given by $\pi_i = 2(-\mu - \sqrt{\Delta_\infty} z + \sqrt{\Delta_0 - \Delta_\infty})$. So that, averaging across the population: $\pi = 2(-\mu + \sqrt{\Delta_0 - \Delta_\infty})$. When poisitive, π returns an overlap; when negative, it measures a distance between the two orbits. Finally, when the two chaotic attractors completely, $\pi = 2\sqrt{\Delta_0}$. We thus define the average overlap as the normalized quantity:

$$\pi = \frac{-\mu + \sqrt{\Delta_0 - \Delta_\infty}}{\sqrt{\Delta_0}} \tag{39}$$

which has a maximum in 1 when the overlap is complete ($\mu = 0$, $\Delta_\infty = 0$).

For every set of the architecture parameters, the theoretical expected value of the overlap can be computed within the DMF framework. In Fig. 10 **b** we show that, in finite-size networks, the transition probability between the two chaotic attractors monotonically increases with the attractors overlap in the phase space.

Appendix E
Non-Gaussian unit rank structures

Our analysis of random networks with unit rank structures relies on a purely statistical mean field description. In Section 7.3 we found that, according to the DMF theory, single neuron activation variables can be approximated by random processes centered around the mean values μ_i (we set for simplicity $I_i = 0$). Because of the χ_{ij} component of the connectivity, which is randomly drawn at every trial, the neural activation variable x_i fluctuates around μ_i with normally distributed deviations of average size $\sqrt{\Delta_0^I}$.

The distribution of mean values μ_i is inherited by the distribution of the structure eigenvector m. When elements m_i are normally distributed, the global probability distribution for the whole network population is given by a convolution of Gaussian distributions, and is thus Gaussian itself (Fig. 11 **a**). In this case, as population averages are written as Gaussian integrals, the DMF equations that we derived are exact.

When m is not a Gaussian vector, the population distribution is in general more complex. In Fig. 11 **b-c**, we show the population distribution obtained from finite networks when the elements in m are uniformly or bimodally distributed. When the random strength g vanishes, the network distribution coincides with the distribution of m. As g is increased, more and more quenched disorder is injected in the network activity by the random connectivity term, and the distribution becomes smoother. When g is sufficiently large, finally, network activity becomes homogeneous: $\mu_i = 0 \ \forall i$ (see paragraph 7.3.3), and the population distribution becomes exactly Gaussian.

At intermediate g values, deriving within the DMF framework the full network distribution requires some additional effort. However, our theoretically predicted DMF values for the first- and second-order momenta can still be adopted to reasonably approximate non-Gaussian distributions. In Fig. 12, we show that indeed the theoretically predicted DMF statistics are in good agreement with the momenta which are measured in finite-size non-Gaussian networks. For the networks we tested, the elements of m are drawn from a uniform or a bimodal distribution. Finally, note that Gaussian descriptions do not trivially extend to other structure distributions when the rank of P_{ij} is larger than one. In those conditions, different distributions can lead to substantially different network dynamics.

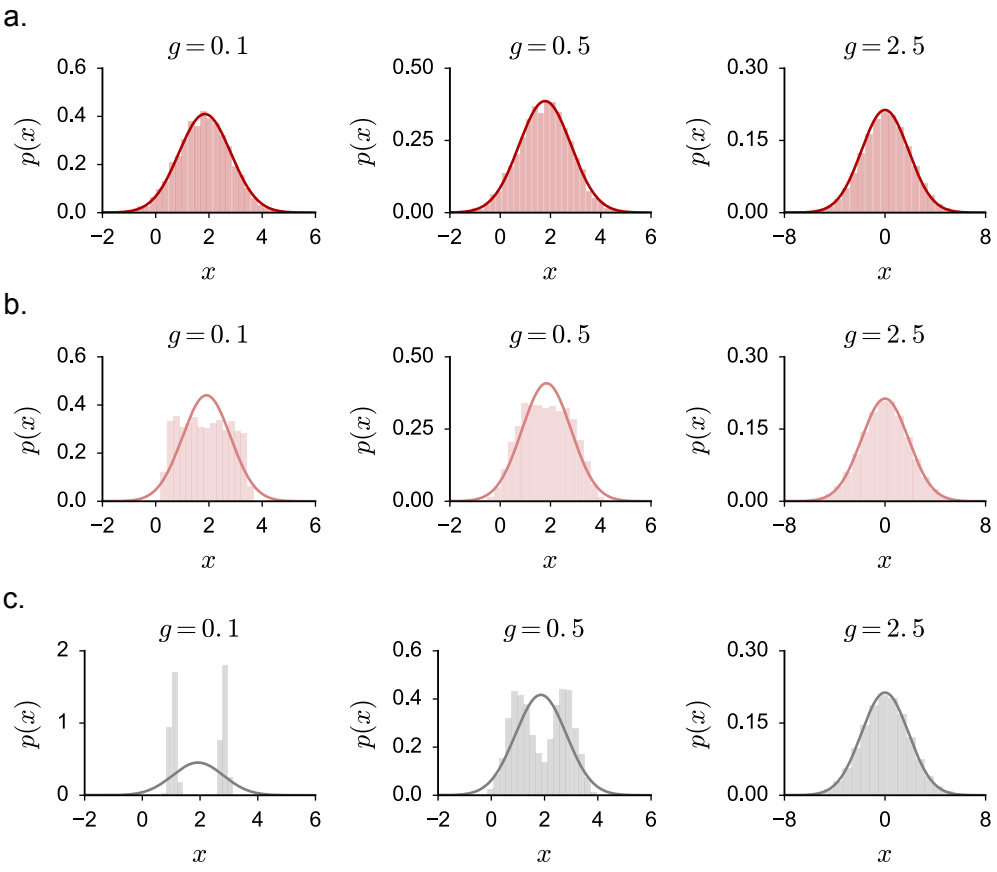


Figure 11: Network activity distribution as the distribution of elements within the structure eigenvector m is varied. Histograms: distribution of numerically simulated activity in a finite-size network ($N = 3000$). Continuous lines: DMF Gaussian prediction. **a.** Elements of m are distributed according to a Gaussian distribution. For any value of g, the population distribution is Gaussian, and is in good agreement with the theoretically predicted probability density. **b-c.** When m is not a Gaussian vector, the population distribution is in general not Gaussian; the agreement with the DMF Gaussian prediction improves as g increases. In **b**, elements of m are drawn randomly from a uniform distribution, centered around M_m; in **c**, the distribution is bimodal: every m_i takes values $M_m - \delta$ and $M_m + \delta$ with equal probability. Choice of the parameters: $\rho = 0$, $M_m M_n = 2.2$.

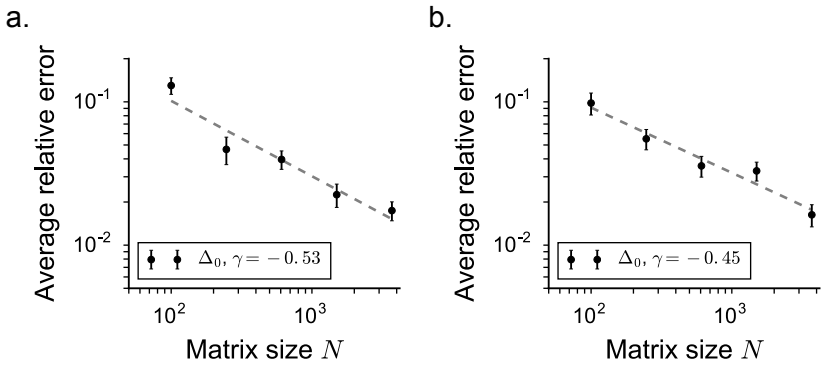


FIGURE 12: Validity of the naive DMF description for network models where the distribution of the structure eigenvector m is not Gaussian. We measure the relative mismatch between the DMF theoretical prediction and the statistics which are numerically measured in finite-size networks as the size of the system is increased. Dashed lines: power-law best fit. Details as in Fig. 7.4 **b**. **a.** Elements of m are drawn randomly from a uniform distribution, centered around M_m. **b.** The distribution of elements of m is bimodal: every m_i takes values $M_m - \delta$ and $M_m + \delta$ with equal probability. Choice of the parameters: $\rho = 0$, $g = 0.6$, $M_m M_n = 2.2$.

Appendix F
Stability analysis in networks with rank two structures

We address here the stability properties of the stationary states which emerge from the dynamics of networks with generic rank two structures.

For those solutions, the expression of the mean field equations for the first- and second-order statistics are determined by the geometrical arrangement of the structure and the input vectors. They have been derived fir different setups in Chapter 8. Similarly to the unit rank case, the simplest mean field solutions correspond to stationary states, which inherit the structure of the most unstable eigenvectors of the connectivity matrix J_{ij}. The stability of the heterogeneous stationary states can be assessed as usual by evaluating separately the value of the radius (Eq. 7.33) and the position of the outliers of the linear stability matrix S_{ij}.

Similarly to the unit rank case, it is possible to compute the position of the outlier eigenvalues by studying the linearized dynamics of the network statistics close to the fixed point, that is given by:

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\begin{pmatrix}\mu^1\\ \Delta_0^1\\ \kappa_1^1\\ \kappa_2^1\end{pmatrix} = -\begin{pmatrix}\mu^1\\ \Delta_0^1\\ \kappa_1^1\\ \kappa_2^1\end{pmatrix} + \mathcal{M}\begin{pmatrix}\mu^1\\ \Delta_0^1\\ \kappa_1^1\\ \kappa_2^1\end{pmatrix}. \tag{40}$$

Note that, in κ_k^l, the subscript $k = 1, 2$ refers to the left vector $n^{(k)}$ with which the overlap is computed, while the superscript $l = 0, 1$ indicates the order of the perturbation away from the fixed point.

In order to compute the elements of the linear stability matrix $\mathcal{M}$, we follow and extend the reasonings that have been discussed in details for the unit rank case. We start by considering the time evolution of the linearized activity μ_i^1, which similarly to Eq. 7.34 reads:

$$\dot{\mu}_i^1(t) = -\mu_i^1 + m_i^{(1)}\kappa_1^1 + m_i^{(2)}\kappa_2^1. \tag{41}$$

At every point in time, we can write: $\mu_i^t = m_i^{(1)}\tilde{\kappa}_1^t + m_i^{(2)}\tilde{\kappa}_2^t$, where $\tilde{\kappa}_k^t$ is the low-pass filtered version of κ_k^t: $(1 + \mathrm{d}\,/\,\mathrm{d}\,t)\tilde{\kappa}_k^t = \kappa_k^t$.

In the case of orthogonal and random structure vectors, we get:

$$\dot{\mu}^1(t) = -\mu^1, \tag{42}$$

so that the elements in the first row of $\mathcal{M}$ vanish. In analogy with Eq. 7.53, the linearized dynamics of Δ_0 gives instead:

$$\dot{\Delta_0^1} = -\Delta_0^1 + 2g^2\langle[\phi_i\phi_i']\rangle\mu^1 + g^2\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\}\Delta_0^1 + 2\Sigma_m^2\kappa_1^0\kappa_1^1 + 2\Sigma_m^2\kappa_2^0\kappa_2^1. \quad (43)$$

Similarly to the unit rank case (Eq. 7.36), in order to determine the linear response of κ_1 we need to compute:

$$\kappa_1^1 = \langle n_i^{(1)}[x_i^1\phi'(x_i^0)]\rangle = \langle n_i^{(1)}\mu_i[\phi_i']\rangle - \left(\frac{\Delta_0^1}{2} - \langle\mu_i^1\mu_i^0\rangle - \langle\mu_i^1\rangle\langle\mu_i^0\rangle\right)\langle n_i^{(1)}[\phi_i'']\rangle \quad (44)$$

A similar expression can be derived for the second first-order statistics κ_2^1.

In general, the integrals in the r.h.s. can be expressed in terms of the perturbations $\tilde{\kappa}_1^1$, $\tilde{\kappa}_2^1$ and Δ_0^1, leading to expressions in the form:

$$\begin{aligned}\kappa_1^1 &= a_{11}\tilde{\kappa}_1^1 + a_{12}\tilde{\kappa}_2^1 + b_1\Delta_0^1\\ \kappa_2^1 &= a_{21}\tilde{\kappa}_1^1 + a_{22}\tilde{\kappa}_2^1 + b_2\Delta_0^1.\end{aligned} \quad (45)$$

Applying the operator $(1 + \mathrm{d}\,/\,\mathrm{d}\,t)$ to the Eq. 44 allows to reshape the results in the final matrix form:

$$\mathcal{M} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2g^2\langle[\phi_i\phi_i']\rangle & g^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} & 2\Sigma_m^2\kappa_1^0 & 2\Sigma_m^2\kappa_2^0 \\ 2b_1g^2\langle[\phi_i\phi_i']\rangle & b_1g^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} & 2b_1\Sigma_m^2\kappa_1^0 + a_{11} & 2b_1\Sigma_m^2\kappa_2^0 + a_{12} \\ 2b_2g^2\langle[\phi_i\phi_i']\rangle & b_2g^2\left\{\langle[\phi_i'^2]\rangle + \langle[\phi_i\phi_i'']\rangle\right\} & 2b_2\Sigma_m^2\kappa_1^0 + a_{21} & 2b_2\Sigma_m^2\kappa_2^0 + a_{22} \end{pmatrix}, \quad (46)$$

The values of the constants a and b depend on the geometric arrangement of the structure and the input vectors. Their value has been computed for different setups throughout Chapter 8.

Bibliography

[1] L. F. Abbott and S. B. Nelson. Synaptic plasticity: taming the beast. *Nat Neurosci*, 2000. URL http://www.nature.com/neuro/journal/v3/n11s/full/nn1100_1178.html.

[2] Y. Abu-Mostafa and J. S. Jacques. Information capacity of the hopfield model. *IEEE Trans. Inf. Theory*, 31(4):461–464, 1985. URL http://ieeexplore.ieee.org/document/1057069/.

[3] Y. Ahmadian, D. B. Rubin, and K. D. Miller. Analysis of the stabilized supralinear network. *Neural Comput.*, 25(8):1994–2037, 2013. URL http://dx.doi.org/10.1162/NECO_a_00472.

[4] D. J. Albers, J. C. Sprott, and W. D. Dechert. Routes to chaos in neural networks with random weights. *Int. J. Bifurc. Chaos*, 08(07):1463–1478, 1998. URL http://www.worldscientific.com/doi/abs/10.1142/S0218127498001121.

[5] J. Aljadeff, D. Renfrew, and M. Stern. Eigenvalues of block structured asymmetric random matrices. *J. Math. Phys.*, 56(10):103502, 2015. URL http://dx.doi.org/10.1063/1.4931476.

[6] J. Aljadeff, M. Stern, and T. Sharpee. Transition to chaos in random networks with cell-type-specific connectivity. *Phys. Rev. Lett.*, 114:088101, 2015. URL http://link.aps.org/doi/10.1103/PhysRevLett.114.088101.

[7] J. Aljadeff, D. Renfrew, M. Vegué, and T. O. Sharpee. Low-dimensional dynamics of structured random networks. *Phys. Rev. E*, 93:022302, 2016. URL http://link.aps.org/doi/10.1103/PhysRevE.93.022302.

[8] S.-I. Amari. Characteristics of random nets of analog neuron-like elements. *IEEE Trans Syst Man Cybern Syst*, SMC-2(5), 1972. URL http://ieeexplore.ieee.org/document/4309193/.

[9] D. J. Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex*, 7(3):237–252, 1997. URL http://www.ncbi.nlm.nih.gov/pubmed/9143444.

[10] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, 1985. URL https://link.aps.org/doi/10.1103/PhysRevLett.55.1530.

[11] M. authors. *Methods and Models in Neurophysics, Lecture Notes of the Les Houches Summer School 2003*. Elsevier Science, 2004.

[12] W. Bair, C. Koch, W. Newsome, and K. Britten. Power spectrum analysis of bursting cells in area mt in the behaving monkey. *J. Neurosci.*, 14(5):2870–2892, 1994. URL http://www.jneurosci.org/content/14/5/2870.

[13] O. Barak. Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.*, 46:1 – 6, 2017. ISSN 0959-4388. URL http://www.sciencedirect.com/science/article/pii/S0959438817300429.

[14] O. Barak, M. Rigotti, and S. Fusi. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *J. Neurosci.*, 33(9):3844–3856, 2013. ISSN 0270-6474. URL http://www.jneurosci.org/content/33/9/3844.

[15] O. Barak, D. Sussillo, R. Romo, M. Tsodyks, and L. Abbott. From fixed points to chaos: Three models of delayed discrimination. *Prog. Neurobiol.*, 103:214 – 222, 2013. URL http://www.sciencedirect.com/science/article/pii/S0301008213000129.

[16] G. Ben Arous and A. Guionnet. Symmetric langevin spin glass dynamics. *Ann. Probab.*, 25(3):1367–1422, 1997. URL http://dx.doi.org/10.1214/aop/1024404517.

[17] R. Ben-Yishai, R. L. Bar-Or, and H. Sompolinsky. Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. USA*, 92(9):3844–3848, 1995. URL http://www.pnas.org/content/92/9/3844.abstract.

[18] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neur. Netw.*, 5(2):157–166, 1994. URL http://ieeexplore.ieee.org/document/279181/.

[19] N. Bertschinger and T. Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Comput.*, 16(7):1413–1436, 2004. URL http://dx.doi.org/10.1162/089976604323057443.

[20] M. Boerlin, C. K. Machens, and S. Denève. Predictive coding of dynamical variables in balanced spiking networks. *PLOS Comput. Biol.*, 9(11):1–16, 2013. URL https://doi.org/10.1371/journal.pcbi.1003258.

[21] K. Britten, M. Shadlen, W. Newsome, and J. Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.*, 12(12):4745–4765, 1992. URL http://www.jneurosci.org/content/12/12/4745.

[22] K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. Responses of neurons in macaque mt to stochastic motion signals. *Visual Neurosci.*, 10(6):1157–1169, 1993. URL https://doi.org/10.1017/S0952523800010269.

[23] C. D. Brody, A. Hernández, A. Zainos, and R. Romo. Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cereb. Cortex*, 13(11):1196–1207, 2003. URL http://dx.doi.org/10.1093/cercor/bhg100.

[24] N. Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci.*, 8(3):183–208, 2000. URL http://dx.doi.org/10.1023/A:1008925309027.

[25] N. Brunel and V. Hakim. Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comput.*, 11(7):1621–1671, 1999. URL http://dx.doi.org/10.1162/089976699300016179.

[26] N. Brunel, F. S. Chance, N. Fourcaud, and L. F. Abbott. Effects of synaptic noise and filtering on the frequency response of spiking neurons. *Phys. Rev. Lett.*, 86:2186–2189, 2001. URL http://link.aps.org/doi/10.1103/PhysRevLett.86.2186.

[27] H. L. Bryant and J. P. Segundo. Spike initiation by transmembrane current: a white-noise analysis. *J. Physiol.*, 260(2):279–314, 1976. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1309092/.

[28] D. V. Buonomano and W. Maass. State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci*, 10(2):113–125, 2009. URL http://dx.doi.org/10.1038/nrn2558.

[29] Y. Burak and I. R. Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLOS Comput. Biol.*, 5(2):1–16, 2009. URL https://doi.org/10.1371/journal.pcbi.1000291.

[30] B. Cessac, B. Doyon, M. Quoy, and M. Samuelides. Mean-field equations, bifurcation map and route to chaos in discrete time neural networks. *Physica D*, 74:24–44, 1994. URL http://www.sciencedirect.com/science/article/pii/0167278994900248.

[31] A. Churchland, R. Kiani, R. Chaudhuri, X.-J. Wang, A. Pouget, and M. Shadlen. Variance as a signature of neural computations during decision making. *Neuron*, 69(4):818 – 831, 2011. URL http://www.sciencedirect.com/science/article/pii/S0896627310010871.

[32] M. M. Churchland and K. V. Shenoy. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J. Neurophysiol.*, 97(6):4235–4257, 2007. URL http://jn.physiology.org/content/97/6/4235.

[33] M. M. Churchland, A. Afshar, and K. V. Shenoy. A central source of movement variability. *Neuron*, 52(6):1085–1096, 2006. URL http://dx.doi.org/10.1016/j.neuron.2006.10.034.

[34] M. M. Churchland, B. M. Yu, S. I. Ryu, G. Santhanam, and K. V. Shenoy. Neural variability in premotor cortex provides a signature of motor preparation. *J. Neurosci.*, 26(14):3697–3712, 2006. URL http://www.jneurosci.org/content/26/14/3697.

[35] M. M. Churchland, B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani, and K. V. Shenoy. Stimulus onset quenches neural variability: a widespread

cortical phenomenon. *Nat. Neurosci.*, 13(3):369–378, 2010. URL http://dx.doi.org/10.1038/nn.2501.

[36] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching. *Nature*, 487 (7405):51–56, 2012. URL http://dx.doi.org/10.1038/nature11129.

[37] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching. *Nature*, 487 (7405):51–56, 2012. URL http://dx.doi.org/10.1038/nature11129.

[38] A. Citri and R. C. Malenka. Synaptic plasticity: Multiple forms, functions, and mechanisms. *Neuropsychopharmacol.*, 33(1):18–41, 2007. URL http://www.nature.com/npp/journal/v33/n1/full/1301559a.html.

[39] A. Compte, N. Brunel, P. Goldman-Rakic, and X.-J. Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex*, 10:910, 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10982751.

[40] A. Crisanti and H. Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Phys. Rev. A*, 36:4922–4939, 1987. URL http://link.aps.org/doi/10.1103/PhysRevA.36.4922.

[41] J. P. Cunningham and B. M. Yu. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.*, 17(11):1500–1509, 2014. URL http://dx.doi.org/10.1038/nn.3776.

[42] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2005.

[43] A. F. Dean. The variability of discharge of simple cells in the cat striate cortex. *Experimental Brain Research*, 44(4):437–440, 1981. URL https://doi.org/10.1007/BF00238837.

[44] G. Deco and E. Hugues. Neural network mechanisms underlying stimulus driven variability reduction. *PLOS Comput. Biol.*, 8(3):1–10, 2012. URL https://doi.org/10.1371/journal.pcbi.1002395.

[45] B. DePasquale, M. M. Churchland, and L. Abbott. Using firing-rate dynamics to train recurrent networks of spiking model neurons. *arXiv preprint*, 2016. URL https://arxiv.org/abs/1601.07620.

[46] B. Doyon, B. Cessac, M. Quoy, and M. Samuelides. Destabilization and route to chaos in neural networks with random connectivity, 1993. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.46.6935&rep=rep1&type=pdf.

[47] B. Dummer, S. Wieland, and B. Lindner. Self-consistent determination of the spike-train power spectrum in a neural network with sparse connectivity. *Front. Comput. Neurosci.*, 8:104, 2014. URL http://journal.frontiersin.org/article/10.3389/fncom.2014.00104.

[48] C. Eliasmith and C. Anderson. *Neural Engineering - Computation, Representation, and Dynamics in Neurobiological Systems*. MIT press, 2004.

[49] M. Fee and J. Goldberg. A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. *Neuroscience*, 198:152 – 170, 2011. URL http://www.sciencedirect.com/science/article/pii/S0306452211011754.

[50] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.*, 61(2):331–349, 1989. URL http://jn.physiology.org/content/61/2/331.

[51] P. Gao and S. Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.*, 32:148–55, 2015. URL https://doi.org/10.1016/j.conb.2015.04.003.

[52] S. Geman and C. R. Hwang. A chaos hypothesis for some large systems of random equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 60(3):291–314, 1982. URL http://dx.doi.org/10.1007/BF00535717.

[53] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014.

[54] V. L. Girko. Circular law. *Theory Probab. Appl.*, 29(4):694–706, 1985. URL http://dx.doi.org/10.1137/1129095.

[55] S. Goedeke, J. Schuecker, and M. Helias. Noise dynamically suppresses chaos in random neural networks. *arXiv preprint*, 2016. URL https://arxiv.org/abs/1603.01880.

[56] R. L. T. Goris, J. A. Movshon, and E. P. Simoncelli. Partitioning neuronal variability. *Nat. Neurosci.*, 17(6):858–865, 2014. URL http://dx.doi.org/10.1038/nn.3711.

[57] A. Grabska-Barwińska and P. E. Latham. How well do mean field theories of spiking quadratic-integrate-and-fire networks work in realistic parameter regimes? *J. Comput. Neurosci.*, 36(3):469–481, 2014. URL http://dx.doi.org/10.1007/s10827-013-0481-5.

[58] O. Harish and D. Hansel. Asynchronous rate chaos in spiking neuronal circuits. *PLOS Comput. Biol.*, 11(7):1–38, 2015. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1004266.

[59] K. D. Harris and T. D. Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58, 2013. URL http://dx.doi.org/10.1038/nature12654.

[60] G. Hennequin, T. P. Vogels, and W. Gerstner. Non-normal amplification in random balanced neuronal networks. *Phys. Rev. E*, 86:011909, 2012. URL http://link.aps.org/doi/10.1103/PhysRevE.86.011909.

[61] G. Hennequin, T. P. Vogels, and W. Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–1406, 2014. URL http://dx.doi.org/10.1016/j.neuron.2014.04.045.

[62] A. Hernandez, V. Nacher, R. Luna, A. Zainos, L. Lemus, M. Alvarez, Y. Vazquez, L. Camarillo, and R. Romo. Decoding a perceptual decision process across cortex. *Neuron*, 66(2):300 – 314, 2010. URL http://www.sciencedirect.com/science/article/pii/S0896627310002345.

[63] N. A. Hessler, A. M. Shirke, and R. Malinow. The probability of transmitter release at a mammalian central synapse. *Nature*, 366(6455):569–572, 1993. URL http://dx.doi.org/10.1038/366569a0.

[64] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4):500–544, 08 1952. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1392413/.

[65] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79(8):2554–2558, 1982. URL http://www.pnas.org/content/79/8/2554.abstract.

[66] C. Huang and B. Doiron. Once upon a (slow) time in the land of recurrent neuronal networks.... *Curr. Opin. Neurobiol.*, 46:31 – 38, 2017. URL http://www.sciencedirect.com/science/article/pii/S0959438817300193.

[67] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks - with an erratum note. *German National Research Center for Information Technology*, 2001. URL http://www.faculty.jacobs-university.de/hjaeger/pubs/EchoStatesTechRep.pdf.

[68] H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004. URL http://science.sciencemag.org/content/304/5667/78.

[69] J. Kadmon and H. Sompolinsky. Transition to chaos in random neuronal networks. *Phys. Rev. X*, 5:041030, 2015. URL http://link.aps.org/doi/10.1103/PhysRevX.5.041030.

[70] P. Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cog. Comput.*, 1(2):139–159, 2009. URL https://doi.org/10.1007/s12559-009-9009-8.

[71] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjostrom, and T. D. Mrsic-Flogel. Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91, 2011. URL http://dx.doi.org/10.1038/nature09880.

[72] H. Ko, L. Cossell, C. Baragli, J. Antolik, C. Clopath, S. B. Hofer, and T. D. Mrsic-Flogel. The emergence of functional microcircuits in visual cortex. *Nature*, 496(7443):96–100, 2013. URL http://dx.doi.org/10.1038/nature12015.

[73] R. Laje and D. V. Buonomano. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nat. Neurosci.*, 16(7):925–933, 2013. URL http://dx.doi.org/10.1038/nn.3405.

[74] E. Ledoux and N. Brunel. Dynamics of networks of excitatory and inhibitory neurons in response to time-dependent inputs. *Front. Comput. Neurosci.*, 5:25, 2011. URL http://journal.frontiersin.org/article/10.3389/fncom.2011.00025.

[75] D. Lee, N. L. Port, W. Kruse, and A. P. Georgopoulos. Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. *J. Neurosci.*, 18(3):1161–1170, 1998. URL http://www.jneurosci.org/content/18/3/1161.

[76] R. Legenstein and W. Maass. Edge of chaos and prediction of computational performance for neural circuit models. *Neural Networks*, 20(3):323 – 334, 2007. URL http://www.sciencedirect.com/science/article/pii/S0893608007000433.

[77] A. Lerchner, G. Sterner, J. Hertz, and M. Ahmadi. Mean field theory for a balanced hypercolumn model of orientation selectivity in primary visual cortex. *Network: Computation in Neural Systems*, 17(2):131–150, 2006. URL http://dx.doi.org/10.1080/09548980500444933.

[78] A. Litwin-Kumar and B. Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat. Neurosci.*, 15(11):1498–1505, 2012. URL http://dx.doi.org/10.1038/nn.3220.

[79] W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.*, 14(11):2531–2560, 2002. URL http://dx.doi.org/10.1162/089976602760407955.

[80] W. Maass, P. Joshi, and E. D. Sontag. Computational aspects of feedback in neural circuits. *PLOS Computat. Biol.*, 3(1):1–20, 2007. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.0020165.

[81] C. K. Machens, R. Romo, and C. D. Brody. Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science*, 307(5712):1121–1124, 2005. URL http://science.sciencemag.org/content/307/5712/1121.

[82] Z. Mainen and T. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506, 1995. URL http://science.sciencemag.org/content/268/5216/1503.

[83] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 11 2013. URL http://dx.doi.org/10.1038/nature12742.

[84] A. Manwani and C. Koch. Detecting and estimating signals in noisy cable structures, i: Neuronal noise sources. *Neural Comput.*, 11(8):1797–1829, 1999. URL http://dx.doi.org/10.1162/089976699300015972.

[85] J. Martens and I. Sutskever. Learning recurrent neural networks with hessian-free optimization. In *ICML*, 2011. URL http://www.icml-2011.org/papers/532_icmlpaper.pdf.

[86] M. Massar and S. Massar. Mean-field theory of echo state networks. *Phys. Rev. E*, 87:042809, 2013. URL http://link.aps.org/doi/10.1103/PhysRevE.87.042809.

[87] F. Mastrogiuseppe and S. Ostojic. Intrinsically-generated fluctuating activity in excitatory-inhibitory networks. *PLOS Comp. Biol.*, 13(4):1–40, 2017. URL https://doi.org/10.1371/journal.pcbi.1005498.

[88] T. Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife*, 6, 2017. URL https://doi.org/10.7554/eLife.20899.

[89] L. Molgedey, J. Schuchhardt, and H. G. Schuster. Suppressing chaos in neural networks by noise. *Phys. Rev. Lett.*, 69:3717–3719, 1992. URL http://link.aps.org/doi/10.1103/PhysRevLett.69.3717.

[90] O. Moynot and M. Samuelides. Large deviations and mean-field theory for asymmetric random recurrent neural networks. *Probab. Theory Relat. Fields*, 123(1):41–75, 2002. URL http://dx.doi.org/10.1007/s004400100182.

[91] B. K. Murphy and K. D. Miller. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009. URL http://dx.doi.org/10.1016/j.neuron.2009.02.005.

[92] J. D. Murray, A. Bernacchia, D. J. Freedman, R. Romo, J. D. Wallis, X. Cai, C. Padoa-Schioppa, T. Pasternak, H. Seo, D. Lee, and X.-J. Wang. A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.*, 17(12):1661–1663, 2014. URL http://dx.doi.org/10.1038/nn.3862.

[93] H. Noda and W. R. Adey. Firing variability in cat association cortex during sleep and wakefulness. *Brain Research*, 18(3):513 – 526, 1970. URL http://www.sciencedirect.com/science/article/pii/0006899370901344.

[94] S. Ostojic. Interspike interval distributions of spiking neurons driven by fluctuating inputs. *J. Neurophysiol.*, 106(1):361–373, 2011. URL http://jn.physiology.org/content/106/1/361.

[95] S. Ostojic. Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons. *Nat Neurosci*, 17(4):594–600, 2014. URL http://dx.doi.org/10.1038/nn.3658.

[96] S. Ostojic and N. Brunel. From spiking neuron models to linear-nonlinear models. *PLOS Comput. Biol.*, 7(1):1–16, 01 2011. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1001056.

[97] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter. Recurrent interactions in spiking networks with arbitrary topology. *Phys. Rev. E*, 85:031916, 2012. URL http://link.aps.org/doi/10.1103/PhysRevE.85.031916.

[98] K. Rajan and L. F. Abbott. Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.*, 97:188104, 2006. URL http://link.aps.org/doi/10.1103/PhysRevLett.97.188104.

[99] K. Rajan, L. F. Abbott, and H. Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Phys. Rev. E*, 82:011903, 2010. URL http://link.aps.org/doi/10.1103/PhysRevE.82.011903.

[100] K. Rajan, C. Harvey, and D. Tank. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128 – 142, 2016. URL http://www.sciencedirect.com/science/article/pii/S0896627316001021.

[101] A. Rauch, G. La Camera, H.-R. Lüscher, W. Senn, and S. Fusi. Neocortical pyramidal cells respond as integrate-and-fire neurons to in vivo–like input currents. *J. Neurophysiol.*, 90(3):1598–1612, 2003. URL http://jn.physiology.org/content/90/3/1598.

[102] A. Renart and C. K. Machens. Variability in neural activity and behavior. *Curr. Opin. Neurobiol.*, 25(Supplement C):211 – 220, 2014. URL http://www.sciencedirect.com/science/article/pii/S0959438814000488.

[103] A. Renart, R. Moreno-Bote, X.-J. Wang, and N. Parga. Mean-driven and fluctuation-driven persistent activity in recurrent networks. *Neural Comput.*, 19(1):1–46, 2006. URL http://dx.doi.org/10.1162/neco.2007.19.1.1.

[104] A. Renart, R. Moreno-Bote, X.-J. Wang, and N. Parga. Mean-driven and fluctuation-driven persistent activity in recurrent networks. *Neural Comput.*, 19(1):1–46, 2007. URL http://dx.doi.org/10.1162/neco.2007.19.1.1.

[105] A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris. The asynchronous state in cortical circuits. *Science*, 327(5965):587–590, 2010. URL http://science.sciencemag.org/content/327/5965/587.

[106] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, 1999.

[107] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013. URL http://dx.doi.org/10.1038/nature12160.

[108] A. Rivkind and O. Barak. Local dynamics in trained recurrent neural networks. *Phys. Rev. Lett.*, 118:258101, 2017. URL https://link.aps.org/doi/10.1103/PhysRevLett.118.258101.

[109] R. Romo, C. D. Brody, A. Hernandez, and L. Lemus. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, 399(6735):470–473, 1999. URL http://dx.doi.org/10.1038/20939.

[110] Y. Roudi and P. E. Latham. A balanced memory network. *PLOS Comput. Biol.*, 3(9):1–22, 2007. URL https://doi.org/10.1371/journal.pcbi.0030141.

[111] D. B. Rubin, S. D. Van Hooser, and K. D. Miller. The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417, 2014. URL http://dx.doi.org/10.1016/j.neuron.2014.12.026.

[112] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. pages 318–362, 1986. URL http://dl.acm.org/citation.cfm?id=104279.104293.

[113] A. Saez, M. Rigotti, S. Ostojic, S. Fusi, and C. D. Salzman. Abstract context representations in primate amygdala and prefrontal cortex. *Neuron*, 87(4):869–881, 2015. URL http://dx.doi.org/10.1016/j.neuron.2015.07.024.

[114] E. S. Schaffer, S. Ostojic, and L. F. Abbott. A complex-valued firing-rate model that approximates the dynamics of spiking networks. *PLOS Comput. Biol.*, 9(10), 2013. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1003301.

[115] P. H. Schiller, B. L. Finlay, and S. F. Volman. Short-term response variability of monkey striate neurons. *Brain Research*, 105(2):347 – 349, 1976. URL http://www.sciencedirect.com/science/article/pii/0006899376904327.

[116] E. Schneidman, B. Freedman, and I. Segev. Ion channel stochasticity may be critical in determining the reliability and precision of spike timing. *Neural Comput.*, 10(7):1679–1703, 1998. ISSN 0899-7667. URL http://dx.doi.org/10.1162/089976698300017089.

[117] T. J. Sejnowski, P. S. Churchland, and J. A. Movshon. Putting big data to good use in neuroscience. *Nat. Neurosci.*, 17(11):1440–1441, 11 2014. URL http://dx.doi.org/10.1038/nn.3839.

[118] H. Seung. How the brain keeps the eyes still. *Proc. Natl. Acad. Sci. USA*, 93(23):13339–13344, 1996. URL http://www.pnas.org/content/93/23/13339.abstract.

[119] M. N. Shadlen and W. T. Newsome. Noise, neural codes and cortical organization. *Curr. Opion. Neurobiol.*, 4 4:569–79, 1994. URL http://www.sciencedirect.com/science/article/pii/0959438894900590.

[120] M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *J. Neurosci.*, 18(10):3870–3896, 1998. URL http://www.jneurosci.org/content/18/10/3870.

[121] N. Shaham and Y. Burak. Slow diffusive dynamics in a chaotic balanced neural network. *PLOS Comput. Biol.*, 13(5):1–26, 2017. URL https://doi.org/10.1371/journal.pcbi.1005505.

[122] M. Shiino and T. Fukai. Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity. *Phys. Rev. E*, 48:867–897, 1993. URL https://link.aps.org/doi/10.1103/PhysRevE.48.867.

[123] O. Shriki, D. Hansel, and H. Sompolinsky. Rate models for conductance-based cortical neuronal networks. *Neural Computat.*, 15(8):1809–1841, 2003. URL http://dx.doi.org/10.1162/08997660360675053.

[124] B. Si, S. Romani, and M. Tsodyks. Continuous attractor network model for conjunctive position-by-velocity tuning of grid cells. *PLOS Comput. Biol.*, 10(4):1–18, 04 2014. URL https://doi.org/10.1371/journal.pcbi.1003558.

[125] A. J. F. Siegert. On the first passage time probability problem. *Phys. Rev.*, 81:617–623, 1951. URL http://link.aps.org/doi/10.1103/PhysRev.81.617.

[126] W. Softky and C. Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *J. Neurosci.*, 13(1):334–350, 1993. URL http://www.jneurosci.org/content/13/1/334.

[127] H. Sompolinsky, A. Crisanti, and H. J. Sommers. Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262, 1988. URL http://link.aps.org/doi/10.1103/PhysRevLett.61.259.

[128] H. F. Song, G. R. Yang, and X.-J. Wang. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, 6:e21492, 2017. URL https://doi.org/10.7554/eLife.21492.

[129] S. Song, P. J. Sjostrom, M. Reigl, S. Nelson, and D. B. Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLOS Biol.*, 3(3), 03 2005. URL https://doi.org/10.1371/journal.pbio.0030068.

[130] M. Stern, H. Sompolinsky, and L. F. Abbott. Dynamics of random neural networks with bistable units. *Phys. Rev. E*, 90:062710, 2014. URL http://link.aps.org/doi/10.1103/PhysRevE.90.062710.

[131] S. Strogatz. *Nonlinear Dynamics And Chaos*. Studies in nonlinearity. Sarat Book House, 2007.

[132] D. Sussillo and L. Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544 – 557, 2009. URL http://www.sciencedirect.com/science/article/pii/S0896627309005479.

[133] D. Sussillo and O. Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computat.*, 25(3):626–649, 2012. URL http://dx.doi.org/10.1162/NECO_a_00409.

[134] D. Sussillo, M. M. Churchland, M. T. Kaufman, and K. V. Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.*, 18 (7):1025–1033, 2015. URL http://dx.doi.org/10.1038/nn.4042.

[135] T. Tao. Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probab. Theory Relat. Fields*, 155(1):231–263, 2013. URL http://dx.doi.org/10.1007/s00440-011-0397-9.

[136] T. Tao, V. Vu, and M. Krishnapur. Random matrices: Universality of esds and the circular law. *Ann. Probab.*, 38(5):2023–2065, 2010. URL http://dx.doi.org/10.1214/10-AOP534.

[137] T. Tetzlaff, M. Helias, G. T. Einevoll, and M. Diesmann. Decorrelation of neural-network activity by inhibitory feedback. *PLOS Computat. Biol.*, 8(8):1–29, 08 2012. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1002596.

[138] D. Thalmeier, M. Uhlmann, H. J. Kappen, and R.-M. Memmesheimer. Learning universal computations with spikes. *PLOS Comput. Biol.*, 12(6):1–29, 2016. URL https://doi.org/10.1371/journal.pcbi.1004895.

[139] B. Tirozzi and M. Tsodyks. Chaos in highly diluted neural networks. *EPL (Europhysics Letters)*, 14(8):727, 1991. URL http://stacks.iop.org/0295-5075/14/i=8/a=001.

[140] G. J. Tomko and D. R. Crapper. Neuronal variability: non-stationary responses to identical visual stimuli. *Brain Research*, 79(3):405 – 418, 1974. URL http://www.sciencedirect.com/science/article/pii/0006899374904387.

[141] T. Toyoizumi and L. F. Abbott. Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. *Phys. Rev. E*, 84:051908, 2011. URL http://link.aps.org/doi/10.1103/PhysRevE.84.051908.

[142] T. W. Troyer and K. D. Miller. Physiological gain leads to high isi variability in a simple model of a cortical regular spiking cell. *Neural Comput.*, 9(5):971—983, 1997. URL http://dx.doi.org/10.1162/neco.1997.9.5.971.

[143] C. van Vreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996. URL http://science.sciencemag.org/content/274/5293/1724.

[144] R. Vogels, W. Spileers, and G. A. Orban. The response variability of striate cortical neurons in the behaving monkey. *Experimental Brain Research*, 77(2):432–436, 1989. URL https://doi.org/10.1007/BF00275002.

[145] C. v. Vreeswijk and H. Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural Comput.*, 10(6):1321–1371, Aug. 1998. URL http://dx.doi.org/10.1162/089976698300017214.

[146] G. Wainrib and J. Touboul. Topological and dynamical complexity of random neural networks. *Phys. Rev. Lett.*, 110:118101, 2013. URL http://link.aps.org/doi/10.1103/PhysRevLett.110.118101.

[147] X.-J. Wang. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):955–968, 2002. URL http://dx.doi.org/10.1016/S0896-6273(02)01092-9.

[148] X.-J. Wang. Decision making in recurrent neuronal circuits. *Neuron*, 60(2):215–234, 2008. URL http://dx.doi.org/10.1016/j.neuron.2008.09.034.

[149] S. Wieland, D. Bernardi, T. Schwalger, and B. Lindner. Slow fluctuations in recurrent networks of spiking neurons. *Phys. Rev. E*, 92:040901, 2015. URL http://link.aps.org/doi/10.1103/PhysRevE.92.040901.

[150] R. C. Williamson, B. R. Cowley, A. Litwin-Kumar, B. Doiron, A. Kohn, M. A. Smith, and B. M. Yu. Scaling properties of dimensionality reduction for neural populations and network models. *PLOS Comput. Biol.*, 12(12):1–27, 12 2016. URL https://doi.org/10.1371/journal.pcbi.1005141.

[151] H. Wilson and J. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, 12:1–24, 1972. URL `https://doi.org/10.1016/S0006-3495(72)86068-5`.

[152] K.-F. Wong and X.-J. Wang. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.*, 26(4):1314–1328, 2006. URL `http://www.jneurosci.org/content/26/4/1314`.

Résumé

Le cortex cérébral des mammifères est constitué de larges et complexes réseaux de neurones. La tâche de ces assemblées de cellules est d'encoder et de traiter, le plus précisément possible, l'information sensorielle issue de notre environnement extérieur. De façon surprenante, les enregistrements électrophysiologiques effectués sur des animaux en comportement ont montré que l'activité corticale est excessivement irrégulière. Les motifs temporels d'activité ainsi que les taux de décharge moyens des cellules varient considérablement d'une expérience à l'autre, et ce malgré des conditions expérimentales soigneusement maintenues à l'identique.

Une hypothèse communément répandue suggère qu'une partie importante de cette variabilité émerge de la connectivité récurrente des réseaux. Cette hypothèse se fonde sur la modélisation des réseaux fortement couplés. Une étude classique [Sompolinsky et al, 1988] a en effet montré qu'un réseau de cellules aux connections aléatoires exhibe une transition de phase: l'activité passe d'un point fixe ou le réseau est inactif, à un régime chaotique, où les taux de décharge des cellules fluctuent au cours du temps et d'une cellule à l'autre. Ces analyses soulèvent néanmoins de nombreuse questions: De telles fluctuations sont-elles encore visibles dans des réseaux corticaux aux architectures plus réalistes? De quelle façon cette variabilité intrinsèque dépend-elle des paramètres biophysiques des cellules et de leurs constantes de temps? Dans quelle mesure de tels réseaux chaotiques peuvent-ils sous-tendre des computations?

Dans cette thèse, on étudiera la dynamique et les propriétés computationnelles de modèles de circuits de neurones à l'activité hétérogène et variable. Pour ce faire, les outils mathématiques proviendront en grande partie des systèmes dynamiques et des matrices aléatoires. Ces approches seront couplées aux méthodes statistiques des champs moyens développées pour la physique des systèmes désordonnées.

Dans la première partie de cette thèse, on étudiera le rôle de nouvelles contraintes biophysiques dans l'apparition d'une activité irrégulière dans des réseaux de neurones aux connections aléatoires. Dans la deuxième et la troisième partie, on analysera les caractéristiques de cette variabilité intrinsèque dans des réseaux partiellement structurées supportant des calculs simples comme la prise de décision ou la création de motifs temporels. Enfin, inspirés des récents progrès dans le domaine de l'apprentissage statistique, nous analyserons l'interaction entre une architecture aléatoire et une structure de basse dimension dans la dynamique des réseaux non-linéaires. Comme nous le verrons, les modèles ainsi obtenus reproduisent naturellement un phénomène communément observé dans des enregistrements électrophysiologiques: une dynamique de population de basse dimension combinée avec représentations neuronales irrégulières, à haute dimension, et mixtes.

Mots Clés

réseaux neuronaux - dynamique des réseaux - théorie du champ moyen - variabilité - neurosciences computationnelles

Abstract

The mammalian cortex consists of large and intricate networks of spiking neurons. The task of these complex recurrent assemblies is to encode and process with high precision the sensory information which flows in from the external environment. Perhaps surprisingly, electrophysiological recordings from behaving animals have pointed out a high degree of irregularity in cortical activity. The patterns of spikes and the average firing rates change dramatically when recorded in different trials, even if the experimental conditions and the encoded sensory stimuli are carefully kept fixed.

One current hypothesis suggests that a substantial fraction of that variability emerges intrinsically because of the recurrent circuitry, as it has been observed in network models of strongly interconnected units. In particular, a classical study [Sompolinsky et al, 1988] has shown that networks of randomly coupled rate units can exhibit a transition from a fixed point, where the network is silent, to chaotic activity, where firing rates fluctuate in time and across units. Such analysis left a large number of questions unsolved: can fluctuating activity be observed in realistic cortical architectures? How does variability depend on the biophysical parameters and time scales? How can reliable information transmission and manipulation be implemented with such a noisy code?

In this thesis, we study the spontaneous dynamics and the computational properties of realistic models of large neural circuits which intrinsically produce highly variable and heterogeneous activity. The mathematical tools of our analysis are inherited from dynamical systems and random matrix theory, and they are combined with the mean field statistical approaches developed for the study of physical disordered systems.

In the first part of the dissertation, we study how strong rate irregularities can emerge in random networks of rate units which obey some among the biophysical constraints that real cortical neurons are subject to. In the second and third part of the dissertation, we investigate how variability is characterized in partially structured models which can support simple computations like pattern generation and decision making. To this aim, inspired by recent advances in networks training techniques, we address how random connectivity and low-dimensional structure interact in the non-linear network dynamics. The network models that we derive naturally capture the ubiquitous experimental observations that the population dynamics is low-dimensional, while neural representations are irregular, high-dimensional and mixed.

Keywords

neural networks - network dynamics - mean field theory - neural variability - computational neuroscience