Stability and amplification in plastic cortical circuits

THÈSE N^{O} 5585 (2013)

PRÉSENTÉE LE 1ER MAI 2013

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE CALCUL NEUROMIMÉTIQUE (IC/SV)
PROGRAMME DOCTORAL EN NEUROSCIENCES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Guillaume HENNEQUIN

acceptée sur proposition du jury:

Prof. C. Petersen, président du jury
Prof. W. Gerstner, directeur de thèse
Prof. M. C. Gastpar, rapporteur
Prof. K. Harris, rapporteur
Prof. W. Senn, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2013

To Maria and François.

Abstract

Cortical circuits are highly recurrently connected, with a globally balanced mix of excitatory (E) and inhibitory (I) synaptic transmission. These E and I interactions among neurons are apparently strong enough to allow the selective amplification of certain patterns of inputs (e.g. sensory inputs from the thalamus) by the collective dynamics of the neurons. Even in the absence of a direct sensory stimulation, amplification is noticeable in the spontaneous formation of non-random activity patterns.

Recurrent amplification raises three fundamental puzzles. First, how do cortical circuits perform such amplification? Second, how can such amplification occur on timescales as fast as those reported during ongoing and stimulus-evoked activities in sensory cortices? Third, how is amplification compatible with the stability of the network dynamics? Strongly recurrent networks are indeed easily prone to dynamical instabilities. This is because some patterns of network activity may be fed back onto themselves by passage through the recurrent connections. This self-feedback may readily be so strong as to cause run-away neuronal activity.

We use a combination of theory and simulations to address these issues. We find the solutions to the three above questions to be one and the same. Stability, amplification, and fast dynamics are simultaneously accounted for in models of cortical circuits in which inhibition is finely tuned to the details of the excitatory pathways. Such networks are shown to exhibit a fine temporal balance between the E and I inputs to single cells, in line with experimental measurements. We also find that the same principles account qualitatively for the dynamics of motor and pre-motor cortical populations during arm-reaching movements in the monkey. Finally, we relate the dynamics of inhibition-stabilized networks to a wider class of dynamical systems known as "nonnormal" in modern physics. This yields a set of theoretical tools with which the behavior of sensory cortical circuits may be conveniently analysed in future studies.

Keywords Cortex, excitation/inhibition balance, nonnormal dynamical systems, stability, amplification, synaptic plasticity.

Résumé

Le cortex forme un réseau de neurones connectés de manière récurrente, où chaque neurone est simultanément excité et inhibé par ses pairs. L'équilibre global entre excitation et inhibition est régulé dynamiquement. Ces interactions sont suffisament fortes pour permettre l'amplification, dans le cortex, de stimuli sensoriels relayés par le thalamus. Une telle amplification se manifeste même par l'apparition spontanée de motifs d'activité corticale en l'absence de stimuli sensoriels.

L'amplification récurrente pose trois énigmes d'un point de vue théorique. Premièrement, par quel mechanisme le cortex est-il capable d'amplifier? Deuxièmement, comment se fait-il que l'amplification se fasse aussi rapidement qu'observé dans les expériences? Enfin, comment se fait-il qu'amplification et stabilité coexistent si harmonieusement? C'est surprenant car les systèmes physiques récurrents sont facilement déstabilisables en général, certains motifs d'activité pouvant être si puissamment renforcés par la dynamique récurrente qu'ils finissent par déstabiliser le système.

Ma thèse approche ces énigmes à travers une combinaison de méthodes théoriques et de simulations sur ordinateur. Il se trouve qu'un seul et même phénomène les explique toutes les trois. Stabilité et amplification sur courte échelle de temps sont simultanément expliquées par des modèles de connectivité corticale dans lesquels les synapses inhibitrices ont été précisément ajustées aux synapses excitatrices. Dans de tels réseaux, les fluctuations temporelles des conductances excitatrices et inhibitrices sont fortement corrélées, en accord avec les mesures expérimentales. Je montre aussi que les mêmes principes semblent gouverner la dynamique collective des neurones du cortex moteur et pré-moteur chez le singe. Enfin, je relie la dynamique des réseaux stabilisés par une inhibition adéquate à une classe plus large de systèmes dits "non-normaux" en physique moderne. Cela me permet de dégager des outils théoriques applicables à l'étude future des circuits corticaux.

Mots clés Cortex, équilibre entre excitation et inhibition, systèmes dynamiques non-normaux, stabilité, amplification, plasticité synaptique.

Acknowledgments

Many people have contributed to this piece of work. Let me thank my PhD advisor Wulfram Gerstner for the (sometimes mind-boggling) freedom he gave me in identifying research topics and pursuing them. I thank him most importantly for persistently issuing the critical words "I think you should now write down what you have" at times when I was about to leave the current project unfinished to explore yet another one[1]. Finally, I thank Wulfram for providing such a nurturing research environment: bringing together the amazing combination of smart colleagues I am about to acknowledge, and making us work together in such a friendly atmosphere, is as much his doing as it is ours.

Thanks to Jean-Pascal Pfister for having patiently taken me through the long delivery process of my first paper. I very much enjoyed his scientific rigor and the many long Skype conversations we had back then.

Very special thanks to Tim Vogels for his enthusiasm about my results; for his constant encouragements; for having channeled my energy into actual papers. Over the past two years, Tim has been the key ingredient in a combination of officemates whose "take-it-easeriously" approach to work has been a breath of fresh air.

Thanks to Chantal Mellier for having so efficiently managed all administrative aspects of the thesis. Having to worry about nothing but packing my toothbrush to go to conference has been such a luxury! I also thank Michèle Bonnard and Sandra Roux for their kindness and patience as I was getting every single step of the thesis submission wrong (or was being late).

Friedemann Zenke has also shaped many of the ideas presented herein. He has also been very helpful in maintaining our computing facilities and reacting fast to my changing needs. I also enjoyed strategizing with Eilif Müller on how to efficiently parallelize my simulations on CPUs as well as on GPUs.

A long list of LCN members must also be acknowledged. Chantal ought to be thanked again for having listed them all up already – the result can be found at http://lcn1.epfl.ch/

[1] Even so, putting my thesis together gave me the disturbing feeling that some of my results will remain buried six feet under – so if you are a PhD student and if you are reading me, it might still be time to listen to your thesis director.

LCNStaff.html, basically from "Jean-Pascal Pfister" all the way up to "Mohammadjavad Faraji". All provided valuable input, scientific and non-scientific. I am particularly grateful to Richard Naud, Danilo Rezende, Alex Seeholzer, Christian Pozzorini, Friedemann Zenke and Henning Sprekeler for having lent critical ears to my mathy wanderings.

I thank the Compagnie Générale de Navigation (CGN, www.cgn.ch) for providing cross-border floating office space, and Nacho Molina for the refreshing discussions onboard.

I wish to thank the large community of open-source developers whose collective work has made my own substantially lighter. In particular, I acknowledge the weekly – if not daily – use of the OCaml programming language (http://caml.inria.fr), the GNU Scientific Library and its OCaml bindings, gnuplot, vim, LaTeX, tikz, Debian GNU Linux, and countless other tools.

Maria, this work is dedicated to you, and is undoubtedly not commensurate with the love and patience you have shown throughout. There would be much more to say than 사랑해요 !

Foreword

The work presented in this thesis is the fruit of five years of research carried out at the Laboratory of Computational Neuroscience (LCN), EPFL, from October 2007 to October 2012.

I started out with a project on Spike Timing-Dependent Plasticity (STDP), a research topic that had been running in the lab since Wulfram Gerstner started it in the mid-1990s. In collaboration with Wulfram Gerstner and Jean-Pascal Pfister, I attempted to bridge two classes of plasticity rules that they had previously developed with Taro Toyoizumi at EPFL. The first class of rules aims at capturing the phenomenology of plasticity with minimal complexity, that is, fitting as much experimental data as possible using simple building blocks. The other approach postulates a functional objective for plasticity and derives the form that activity-dependent synaptic modifications must take to achieve that function. While the first approach does not directly address the functional relevance of synaptic plasticity, the second one threatens to generate mathematical models with no experimental support. The goal of the project was to look for conditions under which both approaches would in fact be the two sides of a same coin. The work led to the publication of a journal article in *Frontiers in Computational Neuroscience*, and appears as chapter 6 in this thesis.

By the end of 2010, my research interests had shifted significantly towards the study of balanced network dynamics. A pair of seminal papers had been published in 2008/9 that introduced the concept of "nonnormal" dynamical systems – of which a theory had originally been developed to explain some phenomena in fluid dynamics – into the field of computational neuroscience. These studies found that nonnormal dynamics could explain two seemingly unrelated cortical phenomena: the fast waxing and waning of non-random activity patterns during ongoing activity in the visual cortex, and the generation of persistent activity in the prefrontal cortex during short-term memory tasks. Picking up on this trend, I explored the extent to which nonnormal effects contribute to the dynamics of randomly connected network, the *de facto* model of microcircuit wiring in theoretical neuroscience. My analysis of nonnormal amplification in random balanced networks has been published in *Physical Review E*, and appears as chapter 2 in this thesis.

Just as I started developing this new research interest, Tim Vogels arrived at the EPFL as a postdoctoral fellow. Tim is an expert in the computational aspects of a physiological

phenomenon called the "detailed excitation/inhibition balance", which denotes the exquisite match between synaptic excitation and inhibition that has been reported fairly recently by a series of experiments. After a few months of collaboration it became clear to us that nonnormal dynamics and the detailed balance were tightly related through inhibitory synaptic plasticity. In chapter 3, I elaborate on this link. In the context of inhibitory-stabilized microcircuits, I establish a relationship between the detailed E/I balance, network stability, and transient amplification, a hallmark of nonnormal dynamical systems. In chapter 4, I further relate optimal stabilization – derived from control-theoretic principles – to the simple inhibitory plasticity rule developed in Vogels et al. (2011). In chapter 5, I study the inhibitory stabilization of cortical networks at a larger, macroscopic scale, and find that a simple and robust solution exists that consists in keeping inhibition local.

Finally, chapter 1 is a general introduction to previous models of cortical dynamics and amplification, and brings up the concepts of nonnormal dynamics, stability, and the excitation-inhibition balance from a mathematical, computational and experimental viewpoint.

Lausanne, October 1st, 2012

G.H.

Contents

CHAPTER 1

Introduction

Isaac Newton famously witnessed the fall of an apple from a tree, which prompted him to write down a set of universal laws to describe the interactions of bodies with mass. These were written in four dimensions, three dimensions for space and one for time. Modern physicists now understand the fall of the apple (and a huge lot of other physics phenomena) with the help of many additional dimensions.

If the fall of the apple is now well understood, we still do not understand the first thing about how Newton and colleagues managed to understand it. Let alone Nobel-winning cognition of that sort, much remains to be understood regarding how brains generate simple behaviour, from simple perception to simple decision-making to simple action-taking. What the brain "does" to generate these multiple facets of behaviour is usually referred to more abstractly as "brain computation". The research presented here hopes to contribute to a scientific field called "computational neuroscience", which seeks to describe the "computations" the brain performs using the same sort of mathematical tools that physicists use to understand how apples fall.

The field has been very successful so far in finding equations that describe how the brain compute at some of the lowest levels, for example how neurons integrate information and communicate with one another. Several excellent textbooks are available on this topic (Dayan and Abbott (2001); Gerstner and Kistler (2002b); Izhikevich (2006), and see also Gerstner and Naud (2009)). Animal and human behaviour being incredibly complex, one naturally

faces the question Newton and colleagues had to answer for their own field: how many dimensions do we need to properly understand brain computations? What can a single neuron achieve? How about two of them? What is gained by making a thousand neurons work together?

A single cubic millimeter of the mammalian neocortex is host to several thousand neurons. Tightly packed, neurons interact in several ways: chemical, molecular, and electrical. This can be considered a hint that isolated neurons cannot do much on their own, and that neurons are wired together for a reason. Computational neuroscience has a long tradition of theoretical work on "neuronal network" dynamics, and the present thesis builds on such theories. Again several neuroscience textbooks are available that summarize over 40 years of research in dynamic network theory (Dayan and Abbott (2001); Gerstner and Kistler (2002b); see also Vogels et al. (2005)).

In section section 1.1 I review the models of network dynamics that are used throughout the thesis. In particular, I highlight the problem of dynamical stability and the mathematical framework in which to study it. I then go on describing the phenomenon of cortical amplification and related models (section 1.2), before introducing "nonnormal" dynamical systems from both mathematical and neuroscientific perspectives (section 1.3). Throughout this introduction, I have tried to limit the information content to only slightly above the minimum prescribed by what the thesis covers. A lot more can be found in the various textbooks mentioned above; www.scholarpedia.org is also a remarkable mine of information.

1.1 Cortical dynamics

1.1.1 Neurons and their time constants

Neurons in the cortex can be seen as tiny electrical devices. As such they have a "voltage" that varies in time and reflects the constant flow (in and out) of charged ions through a cellular membrane that features both resistance and capacitance. Membranes are characterized by a resting voltage, that is, the amount (and types) of ions present inside the cell is dynamically regulated so as to sustain a certain equilibrium potential. Now, imagine injecting into the cell a certain number of positively charged ions (say, sodium ions), infinitely quickly[1]. This will instantaneously elevate the membrane voltage, which then will have to return to its equilibrium value because the membrane wants so[2]. How long is the relaxation process

[1] . . . but gently enough not to physically damage the cell - the two are probably incompatible, but here we only imagine.

[2] Technically, the equilibrium expresses an interplay of forces onto the ions, one due to voltage differences, the other due to concentration differences maintained by ion pumps

going to take? It is going to take about three times a duration that is usually referred to as the "membrane time constant", which is in the $10 - 50$ millisecond range for a typical cortical neuron. It is the product of the membrane's resistance and capacitance. This time constant is denoted by τ in this thesis.

We shall remember two important facts about τ for now. One, that τ is roughly comparable to the speed at which natural sensory stimuli vary in many modalities (vision, tactile sensation, audition, ...). And two, that τ is almost two orders of magnitude smaller than the typical timescale of cognitive processes (several seconds). What these two issues imply for neural processing will be discussed shortly.

Let us already write down an equation for the membrane voltage $V(t)$. Since the membrane has both resistive and capacitive properties (in parallel), we can directly use the voltage differential equation for a standard R-C circuit, which reads

$$\tau \frac{dV}{dt} = -V(t) + V_r + RI(t) \tag{1.1}$$

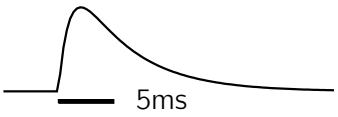
where V_r is the cell's resting potential, R is the resistance of the membrane, and I is the total input current. I comprises all the movements of ions through the membrane that are not already included in the V_r term, as well as the current injected by the experimentalist via an electrode. Equation 1.1 happens to provide a good quantitative description of neuronal responses to input currents, so long as currents are delivered close enough to the soma (main cell body) and do not lead to the generation of "action potentials"[3] (see below). As we shall see later, a large fraction of the inputs is actually delivered quite far away from the soma, a situation for which Equation 1.1 may no longer be accurate. In any event, Equation 1.1 has the key advantage of being simple enough as to allow for a mathematical analysis of network dynamics, as detailed in section 1.2.

Neurons receive inputs from other neurons, mainly through chemical synapses. A synapse is a localized site of close proximity between the membranes of two neurons, usually betwen the axon of the "sender" neuron and a dendrite of the "receiver" neuron. Synapses are equipped with the molecular machinery needed to transmit signals from one side to the next. Classically, the signal from the sender arrives on the pre-synaptic side as a very brief and very large pulse of voltage, which triggers the release of chemicals unsurprisingly called neurotransmitters. These travel to the postsynaptic side where they bind to "receptors". As bindings occurs, the receptors give a certain type of ions a chance to either leave the cell or come in through the membrane, effectively modifying the local membrane conductance. Ions accept or reject the offer, depending on how close the membrane voltage is to their own

[3] If spikes are triggered, Equation 1.1 must be augmented with nonlinear terms, and take into account spike-triggered adaptation currents.

"reversal potential". If ion movement does occur, the net effect on the postsynaptic side will be a depolarizing (excitatory) or hyperpolarizing (inhibitory) current, again depending on the type of ion, which itself depends on the neurotransmitter.

I wish to highlight two important facts about synaptic transmission. First, the window of opportunity given to the ions usually lasts for only a few milliseconds to a few tens of milliseconds, which is at most on the same order as the membrane time constant τ. This yields transient currents with roughly this (absolute) shape:



Second, only a single type of neurotransmitter may be released from the pre-synaptic site, and it is the same for all the synapses formed by the presynaptic partner onto other neurons. In particular, this implies that a neuron cannot excite one cell while simultaneously inhibiting another one. This naturally defines two categories of neurons: excitatory (E) cells and inhibitory (I) cells[4]. We will return to this distinction later on, as it turns out to be critical for network dynamics.

Other molecular processes contribute to shaping the "temporal identity" of single neurons. For example, neurons adapt to their inputs on the 100-500ms timescale, through a dynamic regulation of their "excitability", both intrinsic (e.g. chapter 6) and synaptic ("short-term synaptic plasticity"). The two remarks made above regarding τ apply similarly to these additional time constants: they are both neutral w.r.t. sensory inputs and surprisingly short from the point of view of cognitive tasks.

1.1.2 Neuronal integration: from input to output

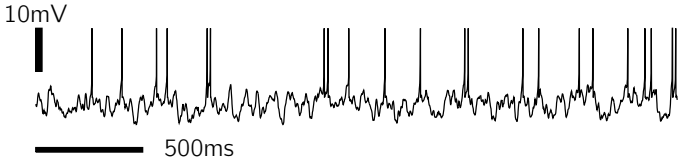
If neurons receive signals from other neurons, it must be that they also sometimes send some. I have introduced synaptic transmission above as being triggered by a "very brief and very large pulse of voltage" produced on the presynaptic side. Such a binary event is called an action potential, or "spike". How does it come about?

As multiple synaptic inputs are being delivered to the cell, their local effects on the postsynaptic voltages are "integrated" into a compound effect at the soma, where their further processing is reasonably well described by Equation 1.1. However, experiments have shown

[4] E and I cells happen to differ in many other respects, e.g. morphology and electrical properties. However, each category embeds sub-categories that also differ in these respects, so the type of neurotransmitter they carry is more defining.

that the input integration itself can be highly nonlinear, depending on the cell's morphology and on where the synapses are placed on the dendritic tree. This is where a drastic approximation is usually made that consists in neglecting any sort of spatial effect altogether. This is called the "point-neuron" approximation: the neuron is thought of as a single point in space, where all synaptic inputs are delivered and merely summed up. One can then continue using Equation 1.1, and $I(t)$ is now the total synaptic input current.

The details on action potential generation are largely irrelevant here. Only two aspects must be mentioned. First, a spike is triggered roughly whenever the cell voltage becomes higher than some threshold. Second, the voltage is approximately reset below threshold following a spike[5]. The spike itself is usually considered a stereo-typical pulse of voltage that is actively generated by the specific transient opening/closing of ion channels, and is usually less than a millisecond long. The spike propagates down the axon to eventually reach all the synapses to which the neuron is a presynaptic partner, which closes the loop. These observations are straightforwardaly expressed in a model of input filtering and spike emission, which produces voltage traces that typically look like this:



Here the input current was taken to be fluctuating with a certain mean and variance, chosen such that the voltage itself fluctuates widely below threshold, yielding occasional action potential firing. This situation is close to the operating regime of cortical networks, as we shall see below.

What we have seen so far is the essence of a family of single-neuron models known as "leaky integrate-and-fire" (LIF) models. Having modeled the behaviour of a single isolated neuron, it is (conceptually) straightforward to carry on and simulate a large pool of such neurons. One only needs to specify their connectivity (and assign values to a great deal of parameters!). This is by now a very standard way of modeling neuronal networks (Vogels et al., 2005). We use just this type of "low-level" modeling in chapter 6, though not in the context of recurrent network dynamics.

Networks of LIF neurons have historically led to a deeper understanding of cortical dynamics. LIF networks can indeed account for various experimentally observed dynamical regimes that range from synchronous firing across the population and rythmicity in single cells, to asynchronous and irregular firing. The latter is thought to be the dynamical regime of cortical microcircuits under "normal" operating conditions. I expand on this regime in the coming

[5] "Roughly" and "approximately" are used on purpose here: the existence of an absolute firing threshold is not clear, and the voltage reset is already a modeling assumption that mimicks the effect of a large and hyperpolarizing spike-triggered intrinsic current.

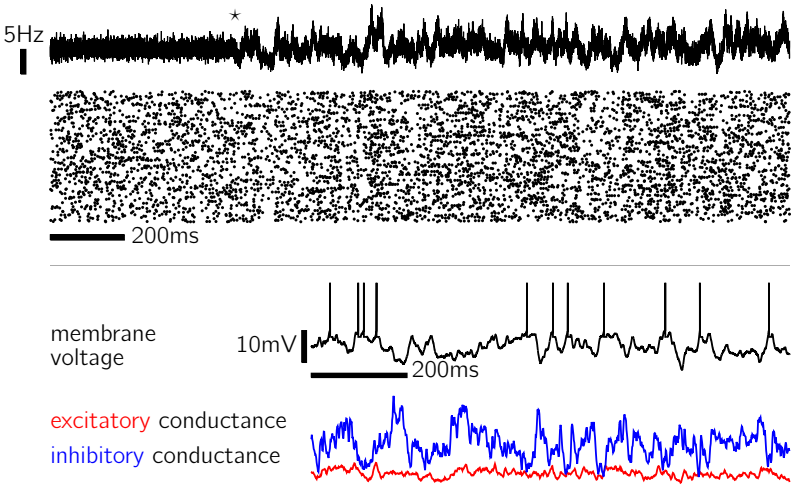


Figure 1.1: **Networks of leaky integrate-and-fire neurons in the balanced state**. A network of 10^5 neurons (80% exc., 20% inh.) with random sparse connectivity generates asynchronous and irregular activity. (**Top**) Timecourse of the momentary overall population firing rate, over 2 seconds. The network was wired at time $t = 500$ms (star). Before wiring, all synaptic inputs to each cell were independent Poisson processes at 5Hz. Most of these artificial inputs were then suddenly replaced by actual network synapses. (**Middle**) Raster plot of single spikes for 500 randomly chosen cells. Each line represents the emission of spikes by one of these cells; dots denote spikes. (**Bottom**) Sample traces of the membrane potential (black), and compound excitatory (red) and inhibitory (blue) input conductances for a single cell.

section, which eventually motivates the reduction of the LIF model to a much simpler form of dynamics that provide the basis for the theoretical analyses of this thesis.

1.1.3 Balanced networks of spiking neurons

As mentioned in subsection 1.1.1, synaptic transmission primarily takes two distinct forms: excitatory and inhibitory. A single cortical neuron receives hundreds of excitatory (E) inputs, which happen to be balanced by hundreds of inhibitory (I) inputs. The E/I balance can be tuned such that the mean compound current and its fluctuations produce subthreshold voltage fluctuations in the postsynaptic cell (Figure 1.1). The net result is a sparse emission of action potentials, which essentially occur whenever the E conductance happens to be greater than average while at the same time the I conductance is lower than average.

As it turns out, such a global balance of E and I inputs can be dynamically regulated in large networks of randomly connected neurons (van Vreeswijk and Sompolinsky, 1996; Brunel, 2000; Vogels et al., 2005; Kumar et al., 2008; Renart et al., 2010). This is illustrated in Figure 1.1. Randomly connected networks can generate asynchronous and irregular spiking

activity with relatively low average firing rates. To understand intuitively how this can be achieved, let us make a simple self-consistency argument (Kumar et al., 2008). Let us imagine that the network is entirely unconnected, and that every neuron receives "artificial" inputs from E and I "shadow" cells, each of which firing irregularly at some constant rate r_0. One may then choose a reasonable value for the strength of the E synapses, and tune the strength of the I synapses so that every network neuron fires irregularly at the very same rate r_0[6]. Now the network cells are statistically indistinguishable from their "shadow" inputs. We may as well replace these shadow inputs by real network neurons, that is, we may wire up the network.

By construction, r_0 is a fixed point of the mean population firing rate in the connected network. Whether it is a stable fixed point depends on several parameters. Stability is easily checked numerically by looking at how a small perturbation Δ of the firing rate r_0 of the shadow cells affects the firing rate of the (unconnected) network neurons. If the network neurons are caused to deviate from r_0 by more than Δ, perturbations of the population firing rate in the connected network are bound to be recurrently amplified, yielding unstable dynamics. In the opposite case, perturbations are suppressed, leading to a stable regime of firing at rate r_0. This situation is made possible if inhibitory feedback is stronger than excitatory feedback. Analytical approaches to the calculation of the steady-state responses exist too for several variants of the leaky integrate-and-fire model (e.g. Richardson (2009); Richardson and Swarbrick (2010)).

Two properties of this balanced state are going to motivate the reduction of the LIF model to a simpler formalism. First, neurons in the balanced state have largely unpredictible spike timings. If their average firing rate is well predicted by theoretical analyses similar in spirit to the self-consistency argument made above[7], the precise times of occurence of single spikes are considerably chaotic (van Vreeswijk and Sompolinsky, 1996). It may therefore make sense to forget about spikes and focus on spike rates instead, which are more reliable quantities. Second, networks of spiking neurons in the balanced state (asynchronous and irregular) behave more linearly than their constituents. Isolated single cells have a highly nonlinear steady-state reponse $r = f(I)$ to a constant input current I. When the cell is embedded in a balanced network, its "f-I" curve smoothens considerably and becomes approximately linear over a broad range of input currents.

[6] The strength of the E synapses is not entirely free though, as it must be large enough to allow for *irregular* firing when considered in conjunction with the I inputs.

[7] Their expected *momentary* probability of emitting a spike may also be characterized as a function of a time-varying input, see e.g. Ledoux and Brunel (2011).

1.1.4 Reduction to rate dynamics

Although the quality of a neuron model depends mainly on its ability to faithfully capture the responses to arbitrary stimuli (Gerstner and Naud (2009)), a good neuron model must also be amenable to a theoretical analysis of how neurons behave collectively. As outlined in the previous section, several methods for analyzing the dynamics of large networks of integrate-and-fire neurons have been developed during the past 25 years (van Vreeswijk and Sompolinsky, 1996; Gerstner, 2000; Brunel, 2000; Kumar et al., 2008; Renart et al., 2010). However, these mean-field techniques usually assume random wiring (but see Lerchner et al. (2006)) and are limited to the description of macroscopic quantities such as the average spike rate of the neurons and the distribution of pairwise correlations. Analysing the phenomena of stability, patterned amplification, and plasticity that this thesis touches upon requires going beyond the mean-field picture.

I now introduce a much simpler model of network dynamics that no longer operates on the level of single spikes, but on spike rates. I am going to use this model in the next two chapters of the thesis, for essentially three reasons. The first one is pragmatic: the reduced model is definitely a good one, according the second sense of "good" mentioned above. Indeed, it considerably eases the analysis of the phenomena I am going to present. Second, making the underlying simplifying assumptions may not be sacrificing too much on the side of biological plausibility. Perhaps the insights we will have gained by using the simplified model may actually extend to more realistic, lower-level models of cortical dynamics. The ultimate check is of course the validation of the results predicted by the simplified model by extensive large-scale simulations of networks of LIF neurons. The third reason is somewhat deeper: chapter 2 and chapter 3 are primarily concerned with the impact of connectivity on network dynamics. In essence, the network connectivity is a set of weighted links that specifies whether any two neurons are connected, in which direction, and how strongly. Such characteristics are naturally summarized in a connectivity matrix, which means that the linear algebra machinery for matrix analysis is going to be useful. The linear version of the model I am about to introduce can be seen as the simplest form of the dynamics that makes the use of such matrix techniques possible.

There are various ways of deriving rate equations from LIF models or similar (it started with Wilson and Cowan (1972)). They all involve figuring out how the total synaptic input current for some neuron i depends on the firing rates of its presynaptic partners, and how neuron i converts the total input current into a firing rate. A simple and intuitive derivation can be found in the appendix of Miller and Fumarola (2011) – but see also Ermentrout (1994); Shriki et al. (2003); Aviel and Gerstner (2006); Ostojic and Brunel (2011). We have seen in subsection 1.1.3 that the balanced state is characterized by fluctuations of single neuron

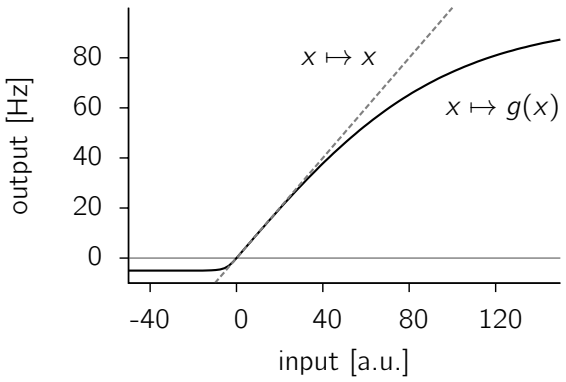
activities $r_i(t)$ around a mean firing rate r_0 (e.g. 5Hz). The rate equation introduced below must be interpreted as an approximate description of these fluctuations *around this mean network activity state*. Let us write the "effective" firing rate of neuron i as the momentary deviation from mean, i.e. $y_i(t) = r_i(t) - r_0$. The rate equation we will use reads

$$\tau \frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{W} g\left[\boldsymbol{x}(t)\right] + \boldsymbol{I}(t) \tag{1.2}$$

Equation 1.2 governs the time-evolution of a vector $\boldsymbol{x}(t) = (x_1(t), x_2(t), \ldots, x_N(t))^T$ of intermediate variables which are related to the effective firing rates $y_i(t)$ through $y_i(t) = g\left[x_i(t)\right]$. Function $g(\cdot)$ reflects (but is not equal to) the "smoooth f-I curve" mentioned in passing at the end of subsection 1.1.3 (see below). Matrix $\boldsymbol{W}$ encodes (mostly) the synaptic connections. τ is a "single-neuron time constant", which roughly compares to the membrane time constant mentioned in subsection 1.1.1, although it also depends on various other parameters such as synaptic timescales. Vector $\boldsymbol{I}(t)$ denotes time-varying inputs that are not a priori part of the network dynamics but which we introduce here in anticipation of later sections.

In the balanced state, the effective impact of neuron j onto neuron i may be either positive or negative, depending on whether neuron j fires momentarily above or below its average r_0. Thus $z \mapsto g(z)$ must assume both negative and positive values. Note also that $g(z)$ cannot be less than $-r_0$, expressing the fact that firing rates cannot be negative. Finally, we must have $g(0) = 0$, because in the limit of small perturbations of the presynaptic partners from their mean rate r_0, the postsynaptic firing rate is equal to r_0 too, that is, $y_i = 0$. Without loss of generality we may also assume $g'(0) = 1$ (the slope may be absorbed in the weight matrix).

When one is not concerned with matching $g(\cdot)$ to a particular spiking neuron model, one may choose g heuristically so long as it satisfies the properties listed above. A reasonable choice could be for example the following curve (see e.g. Rajan et al. (2010)):



The analyses carried out in chapter 2 and chapter 3 further assume a linear gain function in Equation 1.2, yielding the following linear differential equation:

$$\tau \frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{W}\boldsymbol{x}(t) + \boldsymbol{I}(t) \tag{1.3}$$

In this case, $\boldsymbol{x}(t)$ directly encodes the momentary deviations of the firing rates from their means, and we shall refer to $\boldsymbol{x}$ as "firing rate" for simplicity. As we shall see, even a linear model supports interesting phenomena.

1.1.5 Stability

In subsection 1.1.3 we have outlined the notion of stability of the population activity in the spiking network: if one perturbs the momentary overall population rate and let the recurrent dynamics relax, is the perturbation going to grow or to decay? If the perturbation decays, then the ground state is stable: the population as a whole tends to remain in its regime of asynchronous and irregular firing at rate r_0.

The same question of overall stability can be asked in the reduced dynamics of Equation 1.3. The overall population activity is the projection of the vector $\boldsymbol{x}(t)$ of momentary activity onto the uniform pattern $\boldsymbol{u} = (1, 1, 1, \ldots, 1)/N$, where N is the number of neurons in the network. Mathematically, $\mu(t) = \boldsymbol{u}^T \cdot \boldsymbol{x}(t)$. Let us assume that at time $t = 0^-$, we have $\mu(0^-) = 0$. In the spiking network that would mean that the average momentary firing rate across the full network is r_0, though individual variations are allowed. Let us now perturb the network along $\boldsymbol{u}$ at time $t = 0^+$, i.e. add to each neuron the same constant $\Delta > 0$ so that the mean $\mu(t)$ becomes Δ. Right after this positive perturbation, is μ going to decrease or to increase? Let us take the dot product of Equation 1.3 with $\boldsymbol{u}^T$, and receive at time $t = 0^+$

$$\tau \left.\frac{d\mu}{dt}\right|_{t=0^+} = -\mu(0^+) + \boldsymbol{u}^T\boldsymbol{W}\left(\boldsymbol{x}(0^-) + \Delta\boldsymbol{u}\right) \tag{1.4}$$

$$= \Delta\left(\boldsymbol{u}^T\boldsymbol{W}\boldsymbol{u} - 1\right) + \boldsymbol{u}^T\boldsymbol{W}\boldsymbol{x}(0^-) \tag{1.5}$$

The sign of the right-hand side determines the initial reaction of $\mu(t)$ to the perturbation. Now, let us make two simplifying arguments. First, let us remember that we required $\boldsymbol{x}(0^-)$ to be a zero-mean pattern. Since it does not have, a priori, a good reason to correlate with the entries in $\boldsymbol{W}$, and assuming $\boldsymbol{W}$ is a large matrix ($N \gg 1$), we may assume that $\boldsymbol{W}\boldsymbol{x}(0^-)$ is close to zero, and discard it for now. Second, the term $\boldsymbol{W}\boldsymbol{u}$ is a vector in which the i^{th} element sums up all the presynaptic weights of neuron i. A priori, this is not a small vector. However, if the statistics of the presynaptic weights are roughly the same across postsynaptic neurons, then all the entries in $\boldsymbol{W}\boldsymbol{u}$ are roughly equal. This can be written as $\boldsymbol{W}\boldsymbol{u} \simeq \lambda\boldsymbol{u}$

where λ is the expected sum of presynaptic weights per neuron. Thus, Equation 1.5 can be simplified to yield to following stability condition for the overall population activity μ:

$$\lambda < 1 \tag{1.6}$$

We thus expect the fluctuations of the overall population activity around zero to remain stable, provided each neuron receives no more than one unit worth of presynaptic weight. Here the "unit" is arbitrary, but can be related to the units of synaptic conductances in the full spiking model if one derives Equation 1.2 properly. A typical example where this condition is met is a balanced network in which inhibition dominates. In such a network, we even have $\lambda < 0$. Toward the end of chapter 2, we discuss further implications of this global balance condition.

The above arguments, though instructive, are partly incomplete. Indeed, the uniform perturbation $\Delta N\boldsymbol{u}$ is not the only network perturbation that affects μ this way. Any vector that averages to Δ would do. Moreover, one might be able to find such a perturbation $\boldsymbol{u}'$ that is also mapped onto itself through $\boldsymbol{W}$, with a different proportionality constant λ'. The stability condition for $\mu(t)$ would then need to take into account the existence of such a pattern. Finally, we have looked only at the initial growth or decay rate of the perturbation. In principle, even if a perturbation decays initially, it could still end up growing unchecked after a while.

There is a principled way of taking into account all such alternatives. In the language of mathematics, the above observation that $\boldsymbol{W}$ maps vector $\boldsymbol{u}$ onto itself translates into: "$\boldsymbol{u}$ is an eigenvector of $\boldsymbol{W}$". The proportionality constant λ is the associated eigenvalue. Except in very specific cases[8], $\boldsymbol{W}$ has N linearly independent eigenvectors $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N)$ with their associated eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_N)$ (some of them may be identifcal). Thus, the network activity $\boldsymbol{x}(t)$ can at any time be written as a weighted sum of these eigenvectors:

$$\boldsymbol{x}(t) = \sum_{k=1}^{N} a_k(t)\boldsymbol{u}_k \tag{1.7}$$

Note that, in general, the a_k coefficient is not necessarily equal to the projection of $\boldsymbol{x}(t)$ onto $\boldsymbol{u}_k$. The dynamics of the $a_k(t)$ fully decouple in Equation 1.3, yielding

$$\tau\frac{da_k}{dt} = -(1-\lambda_k)a_k(t) \tag{1.8}$$

Here I assume $I(t) = 0$ to focus on the network evolution following some initial condition. The timecourse of each a_k coefficient has a simple solution:

$$a_k(t) = a_k(0)\exp\left(-\frac{t(1-\lambda_k)}{\tau}\right) \tag{1.9}$$

[8] (for defective $\boldsymbol{W}$)

In general, the a_k coefficient as well as the λ_k eigenvalues can be complex numbers. In any case, the norm $\|a_k\|$ decays exponentially following any initial condition, provided the real part of λ_k is smaller than one. In the reverse case, it explodes exponentially.

We may now return to the question of stability of the overall population activity: for $\mu(t)$ to display stable dynamics, the a_k coefficients must decay for all those eigenvectors that have some non-zero overlap with $\boldsymbol{u} \propto (1, 1, \ldots, 1)$. Accordingly, all the corresponding $\mathrm{Re}(\lambda_k)$ must be smaller than unity.

Beyond the overall population activity represented by pattern $\boldsymbol{u} \propto (1, 1, \ldots, 1)$, we may ask whether single neuronal activities are stable too. Let us focus on the first neuron, of which the activity is given by

$$x_1(t) = (1, 0, 0, \ldots, 0)^T \cdot \boldsymbol{x}(t) \tag{1.10}$$

We may expand this according to Equation 1.7:

$$x_1(t) = \sum_{k=1}^{N} a_k(t) \left[(1, 0, \ldots, 0)^T \boldsymbol{u}_k\right] \tag{1.11}$$

A sufficient condition for $x_1(t)$ to remain bounded is that all those eigenvectors of $\boldsymbol{W}$ in which the first entry is nonzero are stable (the $\boldsymbol{u}_k$'s for which the term inside square brackets is non-zero).

"Network stability" in this thesis is defined more broadly as the stability of all possible patterns of activity (equivalent to the stability of all the neurons taken separately). Such general stability is obtained if and only if $\boldsymbol{W}$ has no eigenvalue with real part greater than one.

Important note Linear stability of the network is different from the mere boundedness of the firing rates. Indeed, in the nonlinear rate model of Equation 1.2, the saturating nonlinearity $g(\cdot)$ may automatically prevent run-away activity, but this does not mean the network is "stable" in the linear stability sense. An example of such scenario is given in the next section.

1.1.6 Chaos in rate models of random balanced neuronal networks

How linearly stable are random balanced networks (the focus of chapter 2)? Random balanced networks are broadly defined as networks made of split populations of excitatory and inhibitory neurons, among which connections are drawn at random. Rajan and Abbott (2006) have shown that the eigenvalues of such connectivity matrices are randomly (though not

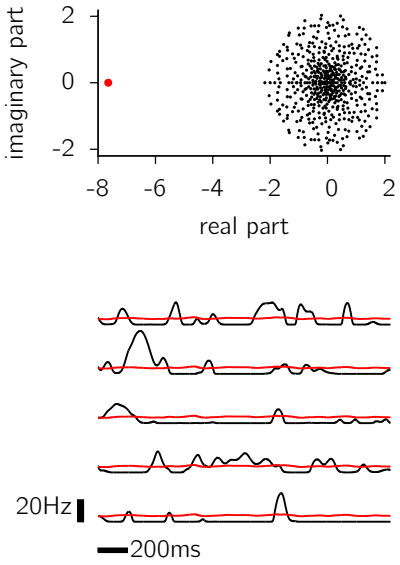


Figure 1.2: **Chaotic fluctuations in linearly unstable balanced neuronal networks**. (**Top**) Eigenvalue spectrum of a random balanced neuronal network, where inhibition is four times as strong as excitation on average. The eigenvalue associated with the "DC" mode is shown in red. (**Bottom**) Five sample single-neuron activity traces. The network was initialized in a random firing state of unit standard deviation. Show here are the first 2 seconds of dynamics according to Equation 1.2 with no additional external input. The average activity across the population (same for all five traces) is shown in red. It fluctuates much less.

necessarily uniformly) scattered inside a disk of radius R centered around the origin in the complex plane (Figure 1.2, black dots, $R = 2$). The radius R depends on the distribution of synaptic efficacies. As we have seen in subsection 1.1.5, there is one additional real eigenvalue associated with the uniform mode of activity $(1, 1, \ldots, 1)$ (Figure 1.2, red dot). This eigenvalue quantifies the absolute difference between the mean excitatory weight and the mean inhibitory weight.

Because the main bulk of eigenvalues is centered around zero, half of them have positive real parts. Thus, for sufficiently strong weights, some of these real parts will exceed one, causing the network to become linearly unstable. When the growth of activity caused by such instabilities is kept in check by the saturation of a nonlinear gain function $g(\cdot)$ (Equation 1.2), an interesting phenomenon emerges: chaotic activity that self-sustains (Sompolinsky et al., 1988; Rajan et al., 2010). Initializing the network in a random state and providing no further external input causes neuronal activities to keep fluctuating asynchronously (Figure 1.2, bottom). Despite the individual fluctuations in the single neurons, the population as a whole shows only little fluctuations (Figure 1.2, bottom, red lines). This is because the overall population activity is associated with the large negative eigenvalue mentioned above.

Chaotic networks have attracted a lot of attention from the 1990s. The firing rate fluctuations that chaotic networks produce (Figure 1.2) closely resemble the temporal patterns of activity that the cortex spontaneously generates. From a functional viewpoint, networks that operate in the weakly chaotic regime (spectral radius not too far above 1) are able to buffer their inputs for periods of time that exceed the single-neuron time constant τ (which I set to $20ms$ in Figure 1.2). In other words, information about an input given at time t can be ex-

tracted from the momentary network state at a time $t + C\tau$ with $C \gg 1$. Thus, on the edge of chaos, a chaotic network can be used as a dynamical substrate for solving complex computations that require some amount of short-term memory (Maass et al., 2002; Bertschinger and Natschläger, 2004; Buonomano and Maass, 2009; Sussillo and Abbott, 2009). Other studies have focused on the possibility of forcing such systems out of chaos with appropriate spatiotemporal patterns of inputs (Molgedey et al., 1992; Sussillo and Abbott, 2009; Rajan et al., 2010), somewhat anticipating the recent discovery that the variability of cortical responses is strongly reduced by the onset of sensory stimuli (Churchland et al., 2010b).

This thesis explores an alternative hypothesis to explain the structure of spontaneous joint firing rate fluctuations in cortical circuits. Here, the large fluctuations observed in ongoing cortical activity are thought to reflect the propagation and amplification of noisy inputs through the recurrent circuitry, and the underlying connectivity matrix is hypothesized to be linearly stable. Unlike the chaos hypothesis, this view explains the spatially structured nature of ongoing activity in sensory cortices, as well as the speed of its fluctuations (Murphy and Miller, 2009). In chapter 3, we link this hypothesis to a few physiological phenomena that so far have not been accounted for in models.

In the coming section, I introduce the phenomenon of cortical amplication and existing models thereof.

1.2 Cortical amplification

Amplification of thalamic inputs by the cortical circuitry is the major motivation for the work presented in chapter 2 and chapter 3. Due to the predominance of cortico-cortical connections over feedforward thalamic inputs, the cortex has long been hypothesized to act as an amplifier of those inputs (Douglas et al., 1995). The cortical representation of sensory stimulus is thought to be dynamically formed by the recurrent dynamics (Fiser et al., 2004). In particular, sharp feature selectivity (a prominent characteristic of sensory neurons in all modalities) can be achieved in models of cortical amplification in which the thalamic input is only weakly tuned (Ben-Yishai et al., 1995; Somers et al., 1995; Sompolinsky and Shapley, 1997; Goldberg et al., 2004).

Even in the absence of an external sensory stimulus, sensory cortices in awake mammals do not remain idle but display ongoing (or "spontaneous") activity fluctuations (Kenet et al., 2003; Fiser et al., 2004; Ferezou et al., 2007; Poulet and Petersen, 2008; Luczak et al., 2009; Gentet et al., 2010; Berkes et al., 2011). In the cat primary visual cortex, those fluctuations occur predominantly along spatial modes of activity that bear a striking resem-

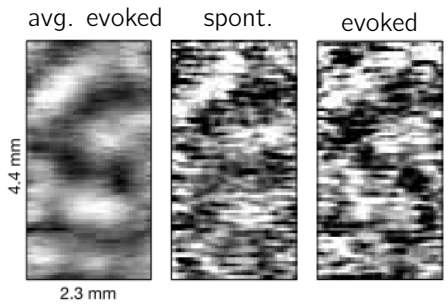


Figure 1.3: **Spontaneous patterned amplification in the visual cortex**. (**Left**) Average map of optically recorded activity in the cat visual cortex, in response to vertically oriented full-field gratings. (**Middle**) Sample activity map (single-frame) during spontaneous activity. (**Right**) Sample activity map (single-frame) of evoked activity, again for full-field vertical gratings. Adapted from Kenet et al. (2003).

blance to sensory-evoked responses (Kenet et al. (2003), Figure 1.3). Similarly, in the rat primary auditory and somatosensory cortices, spontaneous and sensory-evoked activities are statistically similar, both temporally in the order in which neurons tend to fire, and spatially in the joint statistics of spike counts (Luczak et al., 2009).

For the sake of this introduction, let us focus on vision, which as of now is perhaps the sensory modality that has been most theorized about. The data of Kenet et al. (2003) mentioned above are illustrated in Figure 1.3. In the primary visual cortex (V1), neurons are sensitive to the presence of oriented edges in their receptive fields, and each neuron responds preferentially to one given orientation (at least for the so-called "simple cells"). The visual cortex is arranged topologically such that neighbouring neurons have similar preferred orientations. The spatial arrangement of orientation preference can be estimated from optical imaging during sensory-evoked activity. Figure 1.3 (left) shows a typical average pattern of V1 activity in response to vertically oriented gratings. A single frame is shown in Figure 1.3 (right). More surprinsingly, single frames of spontaneous activity could often be observed that resembled such gratings-evoked activity maps (Figure 1.3, middle). The non-randomness of such occurences was significant.

Can network dynamics similar to Equation 1.2 generate patterned amplification, the way V1 seems to be able to?

1.2.1 Amplification by slowing – attractor dynamics

Equation 1.9 on page 11 suggests one way in which the recurrent circuitry could amplify unspecific noisy inputs into specific patterns of network activity. Following some initial condition, each eigenvector $\boldsymbol{u}_k$ of $\boldsymbol{W}$, initially present with some intensity $a_k(0)$, disappears progressively with a time constant $\tau/[1-\mathrm{Re}(\lambda_k)]$ that depends on the corresponding eigenvalue λ_k. The closer the eigenvalue is from instability ($\mathrm{Re}(\lambda_k) \to 1$), the longer the corresponding eigenvector will subsist. Now, imagine that $\boldsymbol{I}(t)$ in Equation 1.3 is a vector of independent

time-varying noisy inputs (say, Gaussian white noise). The network integrates those fluctuating inputs along each eigenvector of $\boldsymbol{W}$, and the intensity of the resulting fluctuations depends on the decay rate imposed by the recurrent dynamics along that eigenmode. For $\mathrm{Re}(\lambda_k) \to 1$, the mode decays slowly so the noise has time to accumulate, yielding slow and large fluctuations of activity pattern $\boldsymbol{u}_k$.

Thus, if one can engineer a connectivity matrix $\boldsymbol{W}$ for which the spatial pattern of evoked V1 activity shown in Figure 1.3 (left) is an eigenvector with large eigenvalue real part, and of which no other eigenvalue is significantly close to 1, then we would expect the spontaneous dynamics to produce activity maps similar to that of Figure 1.3 (middle).

This is the essence of the "ring" model of V1 (Ben-Yishai et al., 1995), and indeed Goldberg et al. (2004) subsequently showed that this mechanism accounts for the data of Kenet et al. (2003) in most respects. A similar principle has been shown recently to account for the dynamics of attention and decision-making in the monkey lateral intraparietal area (Ganguli et al., 2008a).

Amplification-by-slowing makes a strong assumption regarding the speed of the activity fluctuations, both in ongoing and sensory-evoked activities: they must be slow, or at least much slower than the single-neuron time constant τ. However, the data of Kenet et al. (2003) does not appear to show much slowing. Similarly, cortical responses to brief sensory stimuli are often restricted to brief transients.

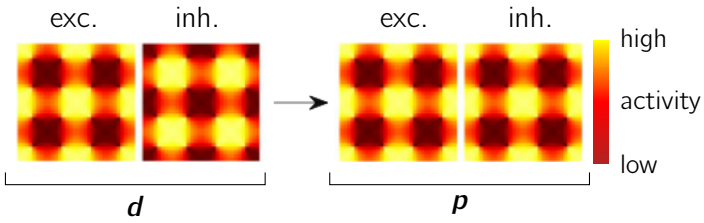
Can amplification be accounted for by a different mechanism that does not rely on slowing?

1.2.2 Transient amplification

The mechanism of amplification-by-slowing, although often complemented by a nonlinear neuronal gain function to turn the spontaneous regime into full attractor dynamics (Goldberg et al., 2004), is essentially a linear mechanism (I used only Equation 1.3 to introduce it). Interestingly, it is not the only linear mechanism that can account for amplification. Murphy and Miller (2009) have indeed provided an alternative linear model of V1 that also explains the data of Kenet et al. (2003) without having to rely on dynamical slowing. It is also more plausible in that is exploits the presence of split populations of excitatory and inhibitory neurons and their balanced interactions.

Amplification somehow must rely on strong interactions among activity patterns. In the case of slowing, the feedback that certain activity modes exert on themselves is strong, and this is what underlies the elongation of the decay rate and amplification thereof. The fundamental observation that Murphy and Miller (2009) made is that strong feedback is not necessary for

amplification, and can be replaced by strong *feedforward* interactions. Along with Murphy and Miller (2009) in the same issue of *Neuron*, Goldman (2009) argued that networks that are fully recurrent in terms of neuronal interactions could in fact be purely feedforward (no feedback) in the way they link activity patterns with one another. In their V1 study, Murphy and Miller (2009) realised that the highly spatially organized *balanced* connectivity of V1, of which they built a model, yields a connectivity matrix $\boldsymbol{W}$ of this feedforward type. In particular, $\boldsymbol{W}$ embeds a strong feedforward link from a certain pattern $\boldsymbol{d}$ to a certain pattern $\boldsymbol{p}$, with the following spatial structures (adapted from Murphy and Miller (2009)):



Here the maps of excitatory (E) and inhibitory (I) network activity are represented separately. This shows that $\boldsymbol{d}$ is a pattern of spatial imbalance (a "difference mode"), in which the maps of E and I activity are spatially opposite to each other. In contrast, $\boldsymbol{p}$ is a spatially balanced pattern of E and I activities (a "sum mode"). Given this feedforward link, how does amplification occur during spontaneous activity in the model? The noisy drive given to each neuron independently provokes random stimulations of pattern $\boldsymbol{d}$, which the recurrent circuitry transforms and amplifies (strong link!) into pattern $\boldsymbol{p}$. As it turns out, $\boldsymbol{p}$ also resembles the map of evoked activity for a specific orientation. Finally, in the model $\boldsymbol{W}$ has no large positive eigenvalue, so that no dynamical slowing would occur.

Both Murphy and Miller (2009) and Goldman (2009) (and one year earlier, Ganguli et al. (2008b)) connected the presence of strong feedforward pathways in $\boldsymbol{W}$ to a more general property of matrices called "nonnormality". I give a short introduction to nonnormal matrices in the following, and the properties of the linear dynamical systems they support.

1.3 Non-normal dynamical systems

To introduce the concept of nonnormality, let us look at a toy two-neuron example network, wired following either of two scenarios:

In the first scenario, neuron x_1 is both excitatory and inhibitory, and so is neuron x_2. In the second scenario, x_1 is purely excitatory, and x_2 is purely inhibitory. The collective dynamics are given by Equation 1.3. How do both systems react to some initial conditions? I constructed both systems to have the same eigenvalue spectrum: in both scenarios the 2×2 connectivity matrix $\boldsymbol{W}$ has two real eigenvalues: $\lambda_u = 0$ and $\lambda_v = -2$. Since none of those eigenvalues are close to 1, we expect no dynamical slowing. It is interesting to note that, for the very same eigenspectrum, synaptic weights in the second scenario assume much larger values than in the first scenario. Figure 1.4 (left) shows the trajectories of $(x_1(t), x_2(t))$ in each scenario, following either of two different initial conditions. In scenario 1, the length of vector $\boldsymbol{x}(t)$ can only decay following any initial condition. We will see why in a moment. In scenario 2, both initial conditions are transiently amplified, as seen from the transient growth of $\|\boldsymbol{x}(t)\|$ (Figure 1.4, right).

The connectivity matrix $\boldsymbol{W}_1$ corresponding to the first scenario is symmetric. A well known property of symmetric matrices is that their eigenvectors are orthogonal to one another. Let

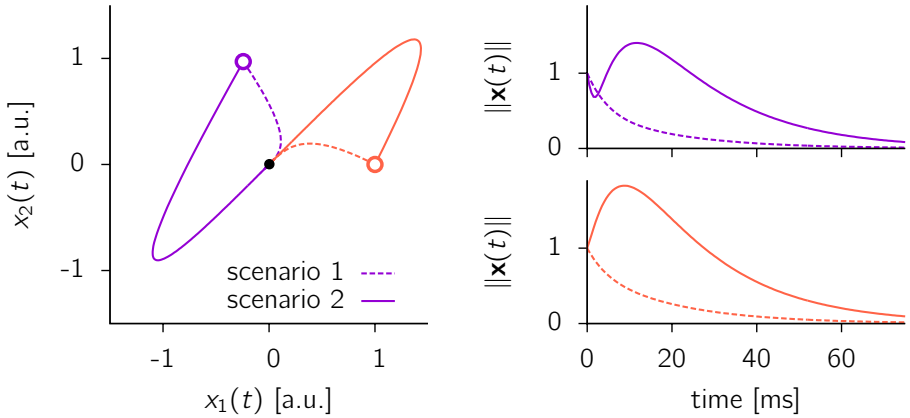


Figure 1.4: **Neural trajectories in our two toy scenarios**. (**Left**) Two initial conditions $(x_1(t = 0), x_2(t = 0))$ (empty circles) were chosen in the space of neuronal activities in our toy two-neuron networks. In scenario 1, the length of the activity vector decays following the initial conditions, returning to rest (black dot) after about 3τ. In scenario 2, activities increase transiently (in absolute magnitude) before returning to rest within about the same duration. (**Right**) Time evolution of the norm $\|\boldsymbol{x}(t)\|$ of the firing rate vector, for the same trajectories.

$\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ denote the two orthogonal eigenvectors of $\boldsymbol{W}_1$. Let us rewrite Equation 1.7 as:

$$\boldsymbol{x}(t) = a_u(t)\boldsymbol{u}_1 + a_v(t)\boldsymbol{v}_1 \tag{1.12}$$

We have seen that, for any matrix $\boldsymbol{W}$, the timecourses of a_u and a_v in absolute values are exponential decays with time constant λ_u and λ_v respectively. In eigenvector coordinates, the squared norm evolves as $\|(a_u(t), a_v(t))\|^2 = a_u^2(t) + a_v^2(t) = \exp(-2t/\lambda_u) + \exp(-2t/\lambda_v)$. Now, because $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are orthogonal, $a_u(t) = \boldsymbol{x}(t)^T \cdot \boldsymbol{u}_1$ and $a_v(t) = \boldsymbol{x}(t)^T \cdot \boldsymbol{v}_1$. Therefore, the squared norm expressed above is *also* the squared norm $\|\boldsymbol{x}(t)\|^2$ of the vector of *firing rates*. Thus, the only way for a network with symmetric connectivity matrix to amplify its inputs is through large positive eigenvalues, so that the rate of the exponential decay slows down. No transient amplification such as the one discovered by Murphy and Miller (2009) can occur.

Symmetry is only a special case of a type of connectivity matrices for which the above also holds true. Any matrix that commutes with its transpose, $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{W}\boldsymbol{W}^T$ has an orthonal eigenbasis, and therefore behaves in the same way as in our first scenario. Any matrix for which the commutation does not hold is called "nonnormal". Such matrices *may* have eigenbases with strongly overlapping eigenvectors, such that the decay of the activity in eigenvector coordinates can hide transient growth of activity in the neurons themselves Murphy and Miller (2009). This is what happens in the second scenario (Figure 1.4). This also relates to the fact that the sum of squares in eigenvector coordinates may have little to do with the sum of squares in the basis of neuronal firing rates (which is the one that is ultimately relevant for amplification).

Important note As the name suggests, "nonnormal" matrices are defined by what they are not: they are not normal. The extent to which transients such as the ones that arise in scenario 2 contribute to the dynamics of a neuronal network depends on *how strongly nonnormal* the connectivity matrix is. Quantifying the degree of nonnormality is not a simple issue, and the paper presented in chapter 2 attempts to achieve precisely such quantification in the case of randomly connected balanced neuronal networks. See also Trefethen and Embree (2005).

Further insights can be obtained from expressing the connectivity matrix $\boldsymbol{W}_2$ in a proper basis of orthogonal activity modes. Matrix $\boldsymbol{W}_2$ is originally expressed in the basis of neurons (vectors $(1, 0)$ and $(0, 1)$). Let us choose another orthogonal basis, made of the following two vectors

$$\boldsymbol{d} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{p} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{1.13}$$

which form the rows of a matrix we call $\boldsymbol{B}$. Note that $\boldsymbol{B}^{-1} = \boldsymbol{B}^T$. In this basis, matrix $\boldsymbol{W}_2$ becomes

$$\boldsymbol{B}\boldsymbol{W}_2\boldsymbol{B}^T = \begin{pmatrix} 0 & 0 \\ 10 & -2 \end{pmatrix} \quad \equiv \quad \boldsymbol{T} \tag{1.14}$$

This matrix describes the functional links between $\boldsymbol{p}$ and $\boldsymbol{d}$ that originally "hide" in $\boldsymbol{W}_2$. For example, $\boldsymbol{p}$ exerts a negative feedback onto itself, with strength -2 (corresponding to λ_v). However, it receives a very strong feedforward link from pattern $\boldsymbol{d}$, with strength 10. The transient activity growth that we saw in scenario 2 originates from this link. The initial condition were in fact purposely chosen roughly along pattern $\boldsymbol{d}$. Interestingly, $\boldsymbol{p}$ does not feed anything back to $\boldsymbol{d}$.

Such a coordinate transform can be generalized to the case of large networks of size N, and is called a "Schur decomposition" (Ganguli et al., 2008b; Goldman, 2009; Murphy and Miller, 2009; Hennequin et al., 2012). In general, a Schur decomposition seeks an orthogonal basis in which $\boldsymbol{W}$ is made triangular, effectively revealing the hidden feedforward connectivity in $\boldsymbol{W}$. The eigenvalues of $\boldsymbol{W}$ line up on the diagonal, and determine the decay rates of each of the Schur modes ($\boldsymbol{d}$ and $\boldsymbol{p}$ here). For normal matrices, importantly, the resulting $\boldsymbol{T}$ matrix is in fact fully diagonal. This means there can be no feedforward links, that is, no transient growth.

The existence of nonnormal matrices has be known for centuries, but it is only recently that the mathematical tools to understand their behaviour have been developed. The primary motivation was to explain some turbulence phenomena in fluid dynamics that could not be explained by standard eigenvalue stability analysis (Trefethen et al., 1993). A book is available that nicely summarizes 20 years of research in related areas of mathematics and physics (Trefethen and Embree, 2005). Of particular interest for this thesis are parts I (introduction to nonnormality and pseudospectra), IV (nonnormal transient effects), VIII (random matrices), IX (computation of pseudospectra) and X (further mathematical issues) of that textbook. In the Supplemental Data of chapter 3 I give further details regarding some of the mathematical tools that help understanding the transient behaviour of nonnormal neural dynamical systems.

1.3.1 Is nonnormality of interest for neural circuits?

The reader may have noticed that the Schur vectors mentioned above have the same spatial structure as the "difference" and "sum" modes mentioned in subsection 1.2.2 (in the V1 model of Murphy and Miller (2009)). In fact, balanced matrices in which columns are either fully positive or fully negative are bound to be nonnormal to some degree Murphy and Miller

(2009), so the dynamics of balanced cortical circuits could in principle be dominated by nonnormal effects. This is definitely the case in the above V1 model.

One of the aims of this thesis is to further understand how such dynamics could contribute to the functioning of neural systems. In chapter 2, we clarify the situation for *random* balanced neuronal networks, a canonical model for cortical microcircuitry at the columnar level. It is found that nonnormality only modestly affect the dynamics. In chapter 3, however, we find that nonnormal dynamics in such microcircuits where inhibition has been finely tuned i) contributes very much and ii) explains further aspects of the cortical physiology that so far have not been explained in other models.

CHAPTER 2

Non-normal amplification in random balanced neuronal networks

This chapter presents the following article:

> Non-normal amplification in random balanced neuronal networks
> **G. Hennequin**, T. P. Vogels and W. Gerstner (2012)
> *Physical Review E*, 86:011909

How nonnormal are random balanced matrices? This question, primarily of a mathematical nature, originates from a problem posed by neuroscience and the study of brain dynamics. In 2009, a series of two papers published simultaneously in *Neuron* have highlighted the importance of the nonnormality of neural connectivity matrices in shaping the collective dynamics of neuronal ensembles. Both studies have been reviewed in depth in chapter 1, but let us recall the findings of interest for this chapter. Murphy and Miller (2009) demonstrated that nonnormal amplification is a major contributor to the macroscopic dynamics of the visual cortex. The authors further argued that, more generally, nonnormality should play a key role in the dynamics of balanced neural networks made of split populations of excitatory and inhibitory cells. Most importantly, nonnormal networks are able to selectively amplify certain patterns of inputs without resorting to traditional slowing mechanisms such as attractor dynamics. The nonnormal model of Murphy and Miller (2009) was thus shown to account for the fast spontaneous fluctuations of the V1 population along non-random activity modes that resemble maps of network activity evoked by oriented visual stimuli. In parallel,

Goldman (2009) showed that the typical patterns of persistent activity routinely observed in prefrontal cortex during short-term memory tasks were also captured by a suitable type of nonnormal dynamics. Goldman argued that networks that look recurrently connected could well be feedforward in disguise, with virtually no feedback connections. Provided the "hidden feedforward connectivity" involves long chains of activity patterns, the network as a whole may generate activity that far outlasts the typically fast time constant τ ($\ll$ 200ms) of its single neurons, i.e. activity that persists for several seconds.

Although each of these two papers brilliantly illustrated their respective points, they did so with rather specific connectivity matrices. One used a model of V1 connectivity (spatially structured and locally dense), the other one used strictly triangular random matrices, which we will see later are definitely abnormally nonnormal, and are not plausible in the sense that neurons may excite some neurons while simultaneously inhibitit some others. These limitations prompted me to look at the nonnormality of random, sparse balanced connectivity matrices, which constitute the *de facto* model of synaptic wiring in cortical microcircuits (Brunel, 2000; Renart et al., 2010). Could nonnormal effects have been overlooked in previous mean-field analyses of such networks?

On the mathematical side, one defining aspect of nonnormal matrices should be recalled here: their eigenvalues do not speak much (Trefethen et al., 1993). Eigenvalue spectra are a common theme in random matrix theory, and the spectral properties of random *balanced* matrices for neuronal networks have already been studied by Rajan and Abbott (2006). However, should it be true that such matrices are significantly nonnormal, the eigenvalues could well show no sign of it. To investigate nonnormal effects on network dynamics, one therefore has to apply a different mathematical framework of which I lay down the bases in this paper.

Let me actually give away the main message. In the end, the results confirm a known result in random matrix theory, namely that random matrices are only weakly nonnormal. This has been shown through several theorems that each used a different approach to quantifying nonnormality (see section 35 in Trefethen and Embree (2005)). By interpreting those matrices as connectivity matrices for neural networks, and by focusing on the impact of their "nonnormal part" on the intensity of spontaneous activity fluctuations, the paper presented in this chapter ends up providing one more theorem of this type. However, the approach we take here allows us to assess the nonnormal contribution to the dynamics in situations where the network would actually be unstable (but not due to nonnormal effects!). Here the findings confirm another series of theorems that demonstrate extreme nonnormality for strictly triangular matrices (section 38 in Trefethen and Embree (2005)). The nonnormal contribution to amplification that we calculate here grows extremely fast indeed in this un-

stable regime. This result turns out to provide one of the main mathematical motivations for studying the inhibitory-stabilization of unstable random networks in the paper presented in chapter 3.

Finally, the results of this paper provide a mechanistic understanding of the chaotic regime of nonlinear random networks with large spectral radii and *asymmetric* (e.g. balanced) connectivity matrices (c.f. chapter 1). In such networks, activity self-sustains (no need for additional network input), and I now believe this is mostly due to the combination of two ingredients. First, unstable random networks embed long chains of strong feedforward interactions among orthogonal activity patterns (the basis for the study of this chapter). And second, they must be simulated with a saturating nonlinearity to prevent runaway activity, and this nonlinearity is likely to incidentally transfer some of the energy accumulated in units that are late in the chain, back to activity modes that are closer to the source. This provides the minimal feedback mechanism for the activity not to die out. If these chains of feedforward interactions are necessary for self-sustained chaotic activity, then normal matrices would not exhibit any. And indeed, simulations of the nonlinear dynamics of a random network of which the connectivity matrix has been symmetrized[1] shows no chaotic activity whatsoever (results not shown). Instead, the network settles in either of a collection of attractor states, depending on the initial condition[2].

There is one more point I would like to clarify as a final comment in this lengthy introduction, as I think it did not come across very clearly in the article. The reader may choose to skip this and return to it after reading the paper, but I believe the following should rather be kept in mind while reading the paper. Throughout the article, we talk about the "normal part" and "nonnormal part" of a matrix $\mathbf{W}$, which we define respectively as the diagonal $\mathbf{\Lambda}$ and the strictly upper-triangular part $\mathbf{T}$ of its Schur decomposition. Does it make sense to "split" a matrix this way, especially when the purpose it to study the nonnormality of $\mathbf{W}$? It should be noted that, although $\mathbf{\Lambda}$ and $\mathbf{T}$ superpose linearly to reconstruct the full Schur triangle, and therefore the connectivity matrix, their respective separate contributions to amplification do *not* add up linearly by any means. In fact, we will see an example in chapter 3 of a matrix in which the eigenvalues interact very strongly with the strict Schur triangle to *reduce* the strength of amplification.

[1] Symmetric matrices are normal matrices

[2] Of course, symmetric matrices are by no means the only type of normal matrices, so to be more conclusive we would need to simulate the dynamics of a much richer set of normal connectivity operators, including e.g. orthogonal networks – which I haven't had time to do.

Abstract

In dynamical models of cortical networks, the recurrent connectivity can amplify the input given to the network in two distinct ways. One is induced by the presence of near-critical eigenvalues in the connectivity matrix $\boldsymbol{W}$, producing large but slow activity fluctuations along the corresponding eigenvectors (dynamical slowing). The other relies on $\boldsymbol{W}$ being nonnormal, which allows the network activity to make large but fast excursions along specific directions. Here we investigate the tradeoff between nonnormal amplification and dynamical slowing in the spontaneous activity of large random neuronal networks composed of excitatory and inhibitory neurons. We use a Schur decomposition of $\boldsymbol{W}$ to separate the two amplification mechanisms. Assuming linear stochastic dynamics, we derive an exact expression for the expected amount of purely nonnormal amplification. We find that amplification is very limited if dynamical slowing must be kept weak. We conclude that, to achieve strong transient amplification with little slowing, the connectivity must be structured. We show that unidirectional connections between neurons of the same type together with reciprocal connections between neurons of different types, allow for amplification already in the fast dynamical regime. Finally, our results also shed light on the differences between balanced networks in which inhibition exactly cancels excitation, and those where inhibition dominates.

2.1 Introduction

A puzzling feature of cortical dynamics is the presence of structure in spontaneously generated activity states. For example, activity in cat primary visual cortex fluctuates along some non-random spatial patterns even when recordings are performed in complete darkness (Tsodyks et al., 1999; Kenet et al., 2003). Similarly, spontaneously generated patterns of firing rates in rat sensory cortices occupy only part of the total space of theoretically possible patterns (Luczak et al., 2009). As the constraints that govern these dynamics cannot be attributed to external stimuli, they are thought to originate from the patterns of synaptic connectivity within the network (Goldberg et al., 2004; Murphy and Miller, 2009). This phenomenon is called patterned amplification.

Patterned amplification can also be observed in simulated neuronal networks, in which spontaneous activity can be modeled as the response to unspecific, noisy inputs delivered to each neuron individually. Propagated through recurrent connections, these noisy inputs may cause the activity of some neurons to transiently deviate from their average more strongly than could be expected from the variability of the external inputs. We thus define amplification here as the strength of these additional, connectivity-induced fluctuations.

Let us consider the following simple linear model for stochastic network dynamics:

$$d\boldsymbol{x} = \frac{dt}{\tau}\left(\boldsymbol{W} - \mathbb{1}\right)\boldsymbol{x} + \sigma_\xi d\boldsymbol{\xi} \tag{2.1}$$

where τ is the neuronal time constant, $\boldsymbol{x} \in \mathbb{R}^N$ is the deviation of momentary network activity with respect to a constant mean firing rate, $\boldsymbol{W}$ is an $N \times N$ synaptic connectivity matrix, $\mathbb{1}$ is the identity matrix, and $d\boldsymbol{\xi}$ is a noise term modeled as a Wiener process. The fluctuations of $x_i(t)$ around zero (i.e. around the mean firing rate of neuron i) are caused by the noisy input and the recurrent drive. Starting from arbitrary initial conditions, the network activity $\boldsymbol{x}$ converges to a stationary Gaussian process with covariance matrix $\boldsymbol{\Sigma} = \left\{\sigma_{ij}\right\}$ (at zero time lag), provided no eigenvalue of $\boldsymbol{W}$ has a real part greater than unity. This covariance matrix has a baseline component $\Sigma_{\text{unc.}} = \sigma_\xi^2 \tau \mathbb{1}/2$ that corresponds to the covariance matrix in the absence of network connections ("unconnected"). Wiring up the network yields additional correlations and potentially gives rise to larger fluctuations of the activity of individual units. We define this amplification A as the ratio $\left[\text{Tr}(\boldsymbol{\Sigma}) - \text{Tr}(\boldsymbol{\Sigma}_{\text{unc.}})\right]/\text{Tr}(\boldsymbol{\Sigma}_{\text{unc.}})$. In other words, A measures the relative gain in mean variance that can be attributed to the recurrent connections. That is,

$$A(\boldsymbol{W}) \quad \stackrel{\text{def}}{=} \quad \left[\frac{2}{\tau\sigma_\xi^2 N}\sum_{i=1}^{N}\sigma_{ii}\right] - 1 \tag{2.2}$$

Under linear dynamics like that of Equation 2.1, amplification can originate from two separate mechanisms. A first, "normal" type of amplification can arise from eigenvalues of $\boldsymbol{W}$ with real parts close to (but smaller than) 1. The noise accumulates along the associated eigenvectors more than in other directions, giving rise to larger activity fluctuations and substantial dynamical slowing along those axes. If the synaptic connectivity is normal in the mathematical sense ($\boldsymbol{W}\boldsymbol{W}^\dagger = \boldsymbol{W}^\dagger\boldsymbol{W}$), it is the *only* mechanism through which the network can amplify its input (Murphy and Miller, 2009). Indeed, if $\boldsymbol{W}$ is normal, its eigenvectors form an orthonormal basis. The sum of variances in this eigenbasis is therefore equal to the sum of variances of the neuronal activities in the original equations. Since linear stability imposes that every eigenvalue of $\boldsymbol{W}$ has a real part less than one, the activity along the eigenvectors can only decay following some initial perturbation. In other words, a stable *normal* linear system is contractive: no initial condition can transiently be amplified. If the matrix $\boldsymbol{W}$ is not normal ($\boldsymbol{W}\boldsymbol{W}^\dagger \neq \boldsymbol{W}^\dagger\boldsymbol{W}$), another, *nonnormal* type of amplification can also contribute (Murphy and Miller, 2009; Ganguli et al., 2008b; Goldman, 2009; Trefethen and Embree, 2005). The eigenvectors are no longer orthogonal to each other, and the apparent decay of the activity in the eigenbasis can hide a transient growth of activity in the neurons themselves. Such growth can only be transient, for stability requirements still demand that the activity decay asymptotically in time.

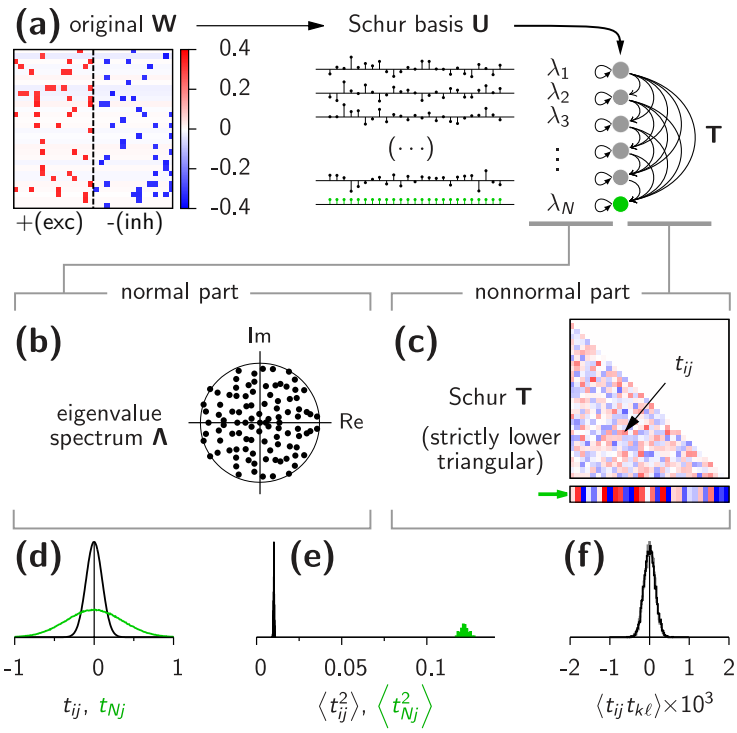


Figure 2.1: **Teasing apart normal and nonnormal amplification in random networks of excitatory and inhibitory neurons**. **(a)** Example sparse neural connectivity matrix $\boldsymbol{W}$ (left, 50 exc. columns and 50 inh. columns, thinned out to 30×30 for better visibility), a schematics of an associated Schur basis $\boldsymbol{U}$ (center), and the corresponding abstract network of Schur modes, in which the interactions are feedforward from top to bottom (right). The Schur vectors in $\boldsymbol{U}$ (center), orthogonal to one another, represent patterns of neuronal activity in the original network. The last Schur vector is explicitly chosen to be the uniform "DC" mode $\boldsymbol{v} = (1, 1, \cdots, 1)/\sqrt{N}$ and is represented here in green at the bottom. **(b)** Amplification via dynamical slowing ("normal" amplification) is described by the set of eigenvalues $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_N)$ of $\boldsymbol{W}$, which for a random network lie inside a disk centered around zero in the complex plane. These eigenvalues determine the decay rates of the Schur patterns. **(c)** Nonnormal amplification arises from the strictly lower-triangular matrix $\boldsymbol{T}$ which describes the purely feedforward part of the interactions between the Schur patterns. The first non-zero entry in the upper left corner of $\boldsymbol{T}$ is t_{21} and represents the "forward" coupling from the first Schur mode onto the second. The last row $(t_{N1}, t_{N2}, \ldots, t_{N(N-1)})$, zoomed-in at the bottom, is the coupling from the first $N-1$ Schur modes onto the last (uniform) Schur mode $\boldsymbol{v}$. **(d)** For a fixed large matrix $\boldsymbol{W}$, the non-zero entries t_{ij} in matrix $\mathbf{T}$ are approximately normally distributed with zero mean and variance given by Equation 2.9 (black (narrow) histogram, for $j < i < N$). The entries in the last row have larger variance given by Equation 2.8 ($i = N$, green (wider) histogram). **(e)** Moreover, the variance $\langle t_{ij}^2 \rangle$ across many realisations of $\boldsymbol{W}$ is the same for all $j < i < N$ (black histogram, left). Similarly, $\langle t_{Nj}^2 \rangle$ is the same for all $j < N$ (green histogram, right). **(f)** The correlations $\langle t_{ij} t_{k\ell} \rangle$ (for $i \neq k$ or $j \neq \ell$) are negligible, as seen from a comparison of their empirical distribution (black) with surrogate data from triangular matrices in which non-zero entries are all drawn i.i.d. (grey, barely visible under the black curve). The data for panels (d–f) was acquired by Schur-transforming 5,000 random weight matrices of size $N = 100$, drawn as described in section 2.3 with connection density $p = 0.1$ and spectral radius $R = 1$.

Purely nonnormal amplification that does not rely on dynamical slowing may be ideally suited for sensory cortices that need to track inputs varying on fast timescales (possibly of order τ). It has also been identified as a key mechanism for short-term memory of past inputs, for in certain circumstances, hidden feedforward dynamics enables the network to retain information about a transient stimulus for a duration of order $N\tau$ (Goldman, 2009). The

presence of noise as in Equation 2.1 could limit this memory duration to $\sqrt{N}\tau$ (Ganguli et al., 2008b; Ganguli and Latham, 2009), but this is still much longer than the time τ in which individual neurons forget their inputs.

The above considerations apply to purposely structured networks (Ganguli et al., 2008b; Goldman, 2009; Murphy and Miller, 2009). It is not clear, however, how much of this beneficial kind of amplification can be expected to arise in randomly connected networks of excitatory and inhibitory neurons, a ubiquitous model of cortical networks. Murphy and Miller (Murphy and Miller, 2009) convincingly argued that nonnormal amplification should generally be a key player in the dynamics of balanced networks, i.e. when strong excitation interacts with equally strong inhibition and when neurons can be only excitatory or inhibitory but not of a mixed type. When the connectivity is dense, or at least locally dense, weak patterns of imbalance between excitation and inhibition can indeed be quickly converted into patterns in which neurons of both types strongly deviate from their mean firing rates. Here, we revisit nonnormal amplification in the context of *random* balanced networks. We derive an analytical expression for the purely nonnormal contribution to amplification in such networks. The analysis reveals a strong tradeoff between amplification and dynamical slowing, suggesting that the connectivity must be appropriately shaped for a network to simultaneously exhibit fast dynamics and patterned amplification.

2.2 Separating the effects of normal and nonnormal amplification

In the Introduction, we distinguished normal from nonnormal amplification. The Schur decomposition (Figure 2.1) – a tool from linear algebra – offers a direct way to assess the contributions of both mechanisms to the overall amount of amplification $A(\boldsymbol{W})$. Any matrix $\boldsymbol{W}$ can be written as $\boldsymbol{U}^\dagger(\boldsymbol{\Lambda}+\boldsymbol{T})\boldsymbol{U}$ where $\boldsymbol{U}=\{u_{ij}\}$ is unitary, $\boldsymbol{\Lambda}$ is a diagonal matrix that contains the eigenvalues λ_k of $\boldsymbol{W}$, and $\boldsymbol{T}=\{t_{ij}\}$ is strictly lower-triangular[3] (Figure 2.1a–c). The lines of $\boldsymbol{U}$ are called the Schur vectors (or Schur modes) and are all orthogonal to each other. If this decomposition is to avoid complex numbers, $\boldsymbol{\Lambda}$ is only block-diagonal, with 2×2 blocks containing the real and imaginary parts of complex conjugate pairs of eigenvalues, and 1×1 blocks containing the real eigenvalues. Importantly, because the Schur basis $\boldsymbol{U}$ is orthonormal, the sum of variances in the basis of the Schur vectors is equal to the sum of the single neuron activity variances. Thus, in order to compute $A(\boldsymbol{W})$, one can instead focus on the activity fluctuations in an abstract network whose units correspond to

[3]Upper, not lower, -triangular $\boldsymbol{T}$ is more common in the literature, but we prefer to keep the flow of information forward (from the 1^{st} to the N^{th} Schur mode) for notational convenience in our calculations.

spatial patterns of neuronal activity (the Schur vectors) and interact with a connectivity matrix $\boldsymbol{\Lambda}+\boldsymbol{T}$ (Figure 2.1a, right). This matrix is lower-triangular, so the abstract network is effectively feedforward. In the Schur network, unit i receives its input from all previous units $j < i$ according to the i^{th} row of $\boldsymbol{T}$. Without input, the activity of unit i decays at a speed governed by eigenvalue λ_i.

A network with a *normal* connectivity matrix would have only self-feedbacks ($\boldsymbol{T} = 0$), thus being equivalent to a set of disconnected units with a variety of individual effective time constants, reflecting dynamical slowing or acceleration. Amplification-by-slowing therefore arises from $\boldsymbol{\Lambda}$ (Figure 2.1b), which summarizes all the "loopiness" found in the original connectivity. Conversely, when $\boldsymbol{\Lambda} = 0$, all units share a common time constant τ (which is also the time constant of the actual neurons) and interact in a purely feedforward manner via matrix $\boldsymbol{T}$ (Figure 2.1c). We refer to this case as "purely nonnormal", because the network is then free of the unique dynamical consequence of normality, namely a modification of the speed of the dynamics[4]. "Purely nonnormal" amplification therefore arises from matrix $\boldsymbol{T}$ that reveals the functional feedforward connectivity hidden in $\boldsymbol{W}$.

The latter situation ($\boldsymbol{\Lambda} = 0$) is the focus of this paper. By substituting $\boldsymbol{W}$ with $\boldsymbol{T}$ in Equation 2.1 and subsequently calculating $A(\boldsymbol{T})$ as defined in Equation 2.2, we intend to reveal the fraction of the total amplification $A(\boldsymbol{W})$ in the neuronal network that cannot be attributed to dynamical slowing, but only to transient growth. This constitutes a functional measure of nonnormality. We carry out this analysis in a statistical sense, by calculating the expected amount of purely nonnormal amplification $\langle A(\boldsymbol{T})\rangle$ where the average $\langle\cdot\rangle$ is over the random matrix $\boldsymbol{W}$. In section 2.3, the ensemble statistics of $\boldsymbol{W}$ are defined, and related to the statistics of the non-zero entries of $\boldsymbol{T}$. In section 2.4 and section 2.5, we perform the calculation of $\langle A(\boldsymbol{T})\rangle$.

2.3 Schur representation of neural connectivity matrices

Prior to calculating the nonnormal contribution to amplification in realistic neural connectivity matrices, we first analyze the statistical properties of the Schur triangle $\boldsymbol{T}$ derived from a neuronal network where every pair of neurons has a certain probability of being connected in either direction. Specifically, we consider networks of $N/2$ excitatory and $N/2$ inhibitory

[4]Note that quantifying nonnormality can be done in a variety of ways, e.g. through several measures of "departure from normality" (Trefethen and Embree, 2005). Our concept of "pure nonnormality" is therefore more specific to our particular purpose, in that it expresses the absence of normal effects on the dynamics of the neurons.

neurons, with connectivity matrices $\boldsymbol{W}$ drawn as follows[5] (Figure 2.1a):

$$w_{ij} = \frac{1}{\sqrt{N}} \cdot \left\{ \begin{array}{ll} +w_0 & \text{if } j \leq N/2 \\ -w_0 & \text{if } j > N/2 \\ 0 \text{ with proba. } & (1-p) \end{array} \right] \text{ with proba. } p \tag{2.3}$$

Excitation and inhibition are thus globally balanced. The $1/\sqrt{N}$ scaling ensures that in the limit of large N, the eigenvalues $\{\lambda_k\}$ of $\boldsymbol{W}$ become uniformly distributed inside the disk of radius

$$R \quad = \quad w_0\sqrt{p(1-p)} \tag{2.4}$$

and centered around zero in the complex plane (Figure 2.1b), with the exception of a few outliers (Rajan and Abbott, 2006). To push the outliers inside the disk, we enforce that excitatory and inhibitory synapses cancel each other precisely for each receiving neuron, i.e. $\boldsymbol{W}\boldsymbol{v} = 0$ with $\boldsymbol{v} = (1, 1, \cdots, 1)/\sqrt{N}$ (Rajan and Abbott, 2006; Tao, 2011). This constraint is also essential to the identification of the ensemble statistics of $\boldsymbol{T}$ as detailed below. Such a "global balance" can be achieved by a Hebbian form of synaptic plasticity at inhibitory synapses in random spiking networks (Vogels et al., 2011). Here we enforce it by subtracting the row average (a small number) from every row (which accounts for the barely visible horizontal stripes in $\boldsymbol{W}$ of Figure 2.1a).

The main point in relating the statistics of $\boldsymbol{T}$ to that of $\boldsymbol{W}$ is to note that the Schur basis is unitary, so that the sum of squares in $\boldsymbol{W}$ is also equal to the sum of squares in $\boldsymbol{\Lambda} + \boldsymbol{T}$. Thus

$$\sum_{1 \leq i,j \leq N} w_{ij}^2 \quad = \quad \sum_{1 \leq k \leq N} |\lambda_k|^2 + \sum_{i>j} t_{ij}^2 \tag{2.5}$$

From our choice of the weights w_{ij} (Equation 2.3) and assuming that N is large enough, we can derive $\sum w_{ij}^2 \simeq Npw_0^2$. Furthermore, knowing that the eigenvalues lie uniformly inside the disk of radius R, we can write $\sum |\lambda_k|^2 \simeq NR^2/2$ which is also valid for large N. We replace these sums in Equation 2.5, simplify the result using Equation 2.4, and obtain the overall empirical variance of the non-zero entries in $\boldsymbol{T}$, to leading order in N:

$$\frac{2}{N(N-1)} \sum_{i>j} t_{ij}^2 \quad \simeq \quad \frac{R^2}{N} \cdot \frac{1+p}{1-p} \tag{2.6}$$

Note that this empirical variance is not necessarily equal to the *ensemble* variance $\left\langle t_{ij}^2 \right\rangle - \left\langle t_{ij} \right\rangle^2$ for fixed i and j. In fact, we have observed that if the non-unique Schur basis is chosen arbitrarily, $\left\langle t_{ij}^2 \right\rangle$ computed over many realisations of $\boldsymbol{W}$ is not uniform across rows, but rather tends to increase with row index i. This heterogeneity is difficult to characterise,

[5] It is straightforward to allow for any distribution of non-zero weights; as it turns out, this Dirac delta distribution achieves maximum nonnormal amplification.

and undermines the calculation of amplification developed in the next section. Fortunately, we can circumvent this problem by choosing the uniform eigenvector $\boldsymbol{v}$ of $\boldsymbol{W}$ as the last Schur vector: $u_{Nk} = 1/\sqrt{N}$ for all k[6]. Coefficient t_{ij} then becomes distributed with the same zero mean and variance ζ^2 for all $j < i < N$, with the exception of the t_{Nj} coefficients which have higher variance ζ_0^2 (black and green lines in Figure 2.1d and Figure 2.1e, empirical observation). Note also that the ensemble pairwise correlations between coupling strengths t_{ij} and $t_{k\ell}$ with $i \neq j$ or $j \neq \ell$ seem negligible (Figure 2.1f).

We now proceed in two steps. First, we focus on the variance of the elements in the *last* row of the Schur matrix $\boldsymbol{T}$, and then we turn to all the other non-zero components. To calculate variance $\zeta_0^2 = \left\langle t_{Nj}^2 \right\rangle$ we use the definition of $\boldsymbol{T}$ and write for $j < N$

$$\begin{aligned} t_{Nj} &= \sum_{\ell=1}^{N}\sum_{k=1}^{N} u_{Nk} w_{k\ell} u_{j\ell} \\ &= \frac{1}{\sqrt{N}} \sum_{\ell=1}^{N} \left(\sum_{k=1}^{N} w_{k\ell} \right) u_{j\ell} \end{aligned} \tag{2.7}$$

To leading order in N we can write $\sum_k w_{k\ell} = \pm p w_0 \sqrt{N}$ where the $\pm$ sign depends on ℓ being smaller than $N/2$ (+, excitatory) or greater (−, inhibitory) – see Figure 2.1a. For $j < N$, the j^{th} Schur vector $\boldsymbol{U}_j$ is orthogonal to the last Schur vector $\boldsymbol{v} \propto (1, 1, \ldots, 1)$, so its components strictly sum to zero: $\sum_\ell u_{j\ell} = 0$. Moreover, because of the normalization, $\sum_\ell u_{j\ell}^2 = 1$. We can therefore approximate $u_{j\ell}$ by a stochastic process with zero mean and variance $1/N$. Assuming the $u_{j\ell}$ are uncorrelated, the variance of t_{Nj} is thus simply $w_0^2 p^2$ to leading order, which according to Equation 2.4 is also

$$\left\langle t_{Nj}^2 \right\rangle \equiv \zeta_0^2 = \frac{R^2 p}{1-p} \tag{2.8}$$

Notably, the variance ζ_0^2 in the last row of coupling matrix $\boldsymbol{T}$ is of order 1, and depends super-linearly on the connectivity density p (Figure 2.2, green lines).

We now turn to the other rows $i < N$ of the Schur matrix $\boldsymbol{T}$. Because all components t_{ij} for $j < i < N$ seem to come from the same distribution and look uncorrelated (Figure 2.1d–f), the empirical estimate of their variance $2\sum_{j<i<N} t_{ij}^2/(N-1)(N-2)$ coincides with the ensemble variance $\zeta^2 \equiv \left\langle t_{Nj}^2 \right\rangle$ so long as N is large enough. Similarly, we can write $\sum_j t_{Nj}^2/(N-1) = \zeta_0^2$. Thus, the l.h.s. of Equation 2.6 becomes $\zeta^2 + 2\zeta_0^2/N$ to leading order in N. Using Equation 2.6 and Equation 2.8 we conclude

$$\left\langle t_{ij}^2 \right\rangle \equiv \zeta^2 = \frac{R^2}{N} \tag{2.9}$$

[6] This is always possible, since a Schur basis can be constructed through Gram-Schmidt orthonormalisation of the eigenbasis of $\boldsymbol{W}$, so choosing $\boldsymbol{v}$ to enter the process first results in $\boldsymbol{v}$ being the last vector in a basis that makes $\boldsymbol{W}$ *lower*-triangular

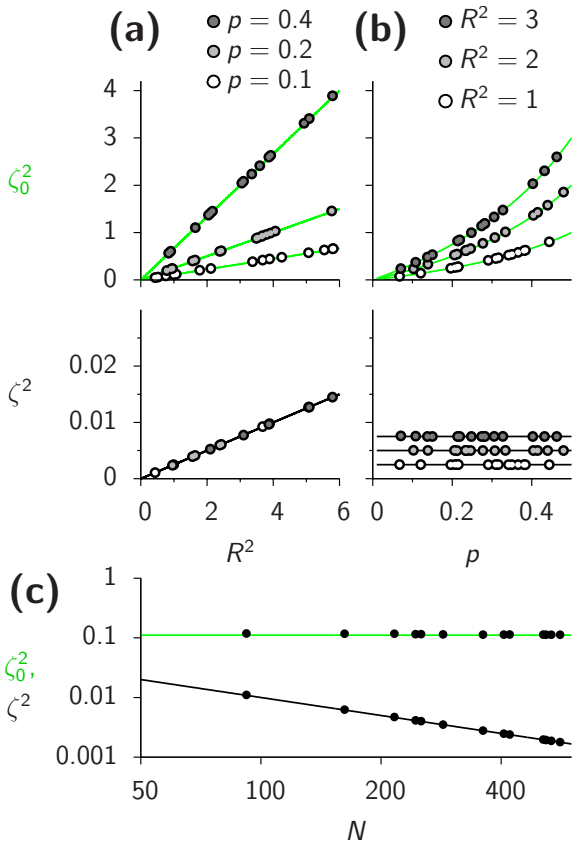


Figure 2.2: **Linking the Schur triangle to the parameters of the neural connectivity matrix.** **(a)** The variance of the entries in the strict lower triangle $\mathbf{T}$ scales linearly with the square of the spectral radius R^2 of the original weight matrix $\mathbf{W}$. For the last row of $\mathbf{T}$, the slope of ζ_0^2 depends on the connection probability p (top plot). For the rest of $\mathbf{T}$, ζ^2 depends only on R^2 (bottom plot). Each point was obtained by empirically estimating ζ^2 and ζ_0^2 from 10 different Schur-transformed random neural weight matrices of size $N = 400$. Lines denote the analytical expressions in Equation 2.8 and Equation 2.9. **(b)** ζ_0^2 in the last row of $\mathbf{T}$ scales super-linearly with the connection density p (top plot). In contrast, ζ^2 does not depend on p (bottom plot). **(c)** In the last row of $\mathbf{T}$, the variance is network size-independent (green (upper) line). In the rest of $\mathbf{T}$, the variance is inversely proportional to N (black (lower) line, note the log-log scale).

Figure 2.2 shows that Equation 2.8 and Equation 2.9 provide a good match to numerical results.

At this point we can already draw a few conclusions. Suppose each unit in our Schur network receives external input of variance 1. First, since the uniform mode $\boldsymbol{v}$ receives network input from the remaining $N-1$ Schur patterns with coupling coefficients of order 1 (Equation 2.8), we expect the global ("DC") population activity $\boldsymbol{x} \cdot \boldsymbol{v}$ to fluctuate macroscopically, i.e. with a variance of order N. In contrast, the rest of the Schur modes should display fluctuations of order 1. Second, we directly see that making the network denser (i.e. increasing p) can only result in larger DC fluctuations, but no further amplification of the other (zero-mean) Schur patterns. This is because ζ_0^2, but not ζ^2, depends on p. Third, it is easy to see where these large DC fluctuations would originate from. Imagine breaking the overall exc.-inh. balance in the network activity by a small amount, e.g. by initialising the network state $\boldsymbol{x}$ to $\boldsymbol{d} = (1, \ldots, 1, -1, \ldots, -1)/\sqrt{N}$, where we emphasize the scaling in $1/\sqrt{N}$. According to Equation 2.1, the transient response to this perturbation is roughly $\boldsymbol{W}\boldsymbol{d}$, which to leading order in N equals

$$\boldsymbol{W}\boldsymbol{d} \quad \simeq \quad pw_0(1, 1, \cdots, 1) \tag{2.10}$$

We note that the $1/\sqrt{N}$ scaling is gone. Thus, the network responds to a microscopic global balance disruption – a state in which the deviation between the excitatory and inhibitory population firing rates is of order $1/\sqrt{N}$ – by an excursion of order 1 in the combined firing rate of both populations (see (Murphy and Miller, 2009) for a more in-depth discussion of this effect). Finally, it is instructive to see what happens when the functional feedforward link from $\boldsymbol{d}$ to $\sqrt{N} \cdot \boldsymbol{v}$ – expressed in Equation 2.10 – is removed from $\boldsymbol{W}$. This can be achieved by transforming $\boldsymbol{W}$ into $\boldsymbol{W}'$ given by

$$\boldsymbol{W}' = \boldsymbol{W} - \frac{pw_0}{\sqrt{N}}(1, \cdots, 1)^\dagger(1, \cdots, 1, -1, \cdots, -1) \tag{2.11}$$

It is easy to see that $\boldsymbol{W}'\boldsymbol{d} = 0$. In this case, calculations similar to Equation 2.5–Equation 2.8 yield $\zeta_0^2 = \zeta^2 = R^2/N$ so that the DC fluctuations are back to order 1: the amplification along the DC mode becomes comparable in magnitude to the amplification that occurs along any other Schur directions. Note that the operation in Equation 2.11 effectively shifts the mean excitatory (resp. inhibitory) weight from $pw_0/\sqrt{N}$ (resp. $-pw_0/\sqrt{N}$) to zero. We now substantiate these preliminary conclusions through a direct calculation of nonnormal amplification.

2.4 Amplification in random strictly triangular networks

We have seen in the preceding two sections that a randomly coupled network of excitatory and inhibitory neurons can be transformed via a unitary Schur basis into a different network where the couplings between units are given by a lower triangular matrix (Figure 2.1a). Furthermore, the "purely nonnormal" part of the amplification of the external noisy input in the original network of neurons corresponds to the activity fluctuations in the new feedforward network where all self-couplings are neglected (Figure 2.1c). Finally, we have also seen that it is possible to constrain the Schur basis such that the couplings between the first $N-1$ units in the feedforward network are independently distributed with the same zero-mean and a variance given by the parameters of the original synaptic weights (Equation 2.9). In this section, we therefore study this "canonical" case, starting directly from a strictly lower-triangular matrix $\boldsymbol{T}$ and ignoring – for the moment – the transformation that gave rise to $\boldsymbol{T}$.

We want to solve for the expected variances of $N \gg 1$ Ornstein-Uhlenbeck processes (as in Equation 2.1) coupled by a strictly lower-triangular weight matrix $\boldsymbol{T}$ (therefore describing a purely feedforward network, see inset in Figure 2.3a). We assume all non-zero coupling strengths to be sampled i.i.d. from some common distribution with zero mean and variance α^2/N. Due to the coupling matrix, the fluctuations that the external input causes in the

first unit feed and augment those it causes in unit 2. The third unit in turn fluctuates due to the external input and the activities of units 1 and 2, and so on. We therefore expect the activity variance σ_{ii} in unit i to increase with index i. In appendix section 2.A, we show that in the limit of large N and for some fixed $0 \leq x \leq 1$, the relative expected variance of the activity in unit $i = xN$ is $g(i/N) \equiv 2\langle\sigma_{ii}\rangle / \tau\sigma_\xi^2$ where the function $g(x)$ is lower-bounded in closed form by

$$\begin{aligned} g^{\mathrm{LB}}(x) &= \frac{1}{3+\sqrt{3}} \exp\left(\frac{1-\sqrt{3}}{4}\alpha^2 x\right) \\ &+ \frac{2+\sqrt{3}}{3+\sqrt{3}} \exp\left(\frac{1+\sqrt{3}}{4}\alpha^2 x\right) \end{aligned} \tag{2.12}$$

(Figure 2.3, dashed blue curves). We also derive the exact solution as a power series

$$g(x) = \lim_{K\to\infty} \sum_{k=0}^{K} \beta_k x^k \tag{2.13}$$

with the β_k coefficients defined recursively as

$$\begin{aligned} \beta_0 &= 1 \\ \beta_k &= \frac{\alpha^2}{2k!} \sum_{\ell=0}^{k-1} \frac{(2\ell)!\,(k-\ell-1)!}{\ell!\,(\ell+1)!} \left(\frac{\alpha^2}{4}\right)^\ell \beta_{k-\ell-1} \end{aligned} \tag{2.14}$$

The overall amplification $A_0(\alpha^2)$ in the network is subsequently obtained by integrating this variance profile $g(x)$ from 0 to 1, which corresponds to taking Equation 2.2 to its $N \to \infty$ limit:

$$A_0(\alpha^2) = \left(\lim_{K\to\infty} \sum_{k=0}^{K} \frac{\beta_k}{k+1}\right) - 1 \tag{2.15}$$

Figure 2.3 shows that Equation 2.13 and Equation 2.15 indeed converge to the empirical mean variance profile and mean amplification as the cut-off parameter K of the power series becomes large (red lines, $K = 10$). Figure 2.3b furthermore shows how amplification explodes with the variance α^2/N of the feedforward couplings in the network.

2.5 Amplification in random balanced networks

Using the canonical result of the previous section that is restricted to homogeneous random lower-triangular matrices, we now calculate $A(R, p) \equiv \langle A(\boldsymbol{T})\rangle$ with $\boldsymbol{T}$ originating from the Schur decomposition of a neuronal connectivity matrix as in section 2.3, with connection density p and spectral radius R. Equation 2.13 can directly be applied with $\alpha^2/N = \zeta^2 = R^2/N$ (see Equation 2.9) to describe the activity fluctuations of the first $N-1$ Schur modes.

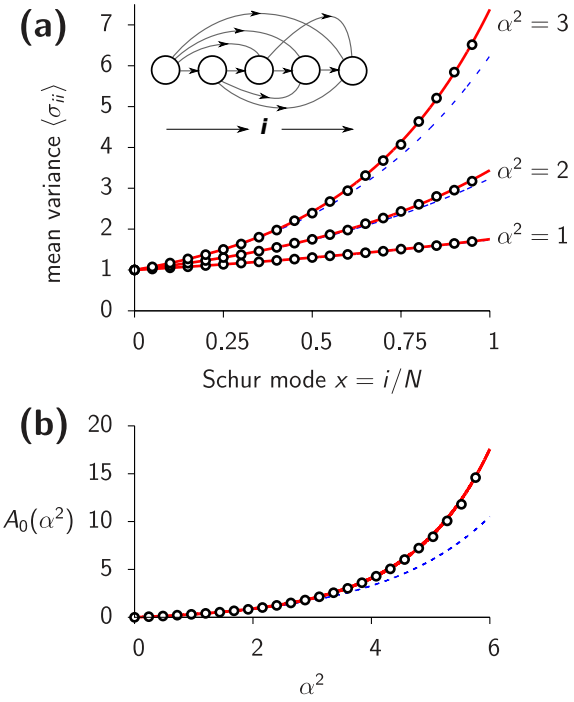


Figure 2.3: **Analytical result for a feedforward network of N Ornstein-Uhlenbeck processes coupled via a random strictly lower-triangular matrix (inset).** **(a)** The expected activity variance $\langle \sigma_{ii} \rangle$ accumulates super-linearly from the first unit to the last down the feedforward chain. Dashed blue lines depict the closed-form lower-bound of Equation 2.12. Solid red lines denote the exact solution given in Equation 2.13, truncated to $K = 10$. Open circles represent the numerical solution of Equation 2.1 – or more exactly, the numerical solution of *Equation* 2.20 given in the appendix – averaged over 20 randomly generated matrices of size $N = 500$. Each matrix $\boldsymbol{T}$ is characterised by the variance α^2/N of the coupling coefficients t_{ij} with $j < i$. The strength of the external noise driving each unit independently is set to $\sigma_\xi^2 = 2/\tau$ so that all activity variances in the network would be 1 should the couplings t_{ij} be set to 0. **(b)** The total amplification (the area under the curves in (a), minus 1) explodes with increasing variance α^2/N in the triangular connectivity matrix. Points and lines have the same meaning as in (a).

The last Schur unit, however, receives feedforward input with couplings of variance $\zeta_0^2 \neq \zeta^2$ (Equation 2.8). Consequently, the expected variance $\langle \sigma_{NN} \rangle$ of its temporal fluctuations has to be treated separately. In appendix section 2.B, we show that

$$\lim_{N\to\infty} \frac{\langle \sigma_{NN} \rangle}{N} \quad = \quad \frac{\sigma_\xi^2 \tau}{2} \cdot \frac{p}{1-p} \left[g\,(1) - 1 \right] \tag{2.16}$$

where g is given by Equation 2.13 and Equation 2.14, here with $\alpha = R$. Gathering the contributions of all Schur modes, we obtain the expected overall amount of purely nonnormal amplification in $\boldsymbol{W}$:

$$A(R,p) \quad = \quad A_0(R^2) + \frac{p}{1-p} \left[g\,(1) - 1 \right] \tag{2.17}$$

with $A_0(R^2)$ given by Equation 2.15.

Figure 2.4a shows that the nonnormal contribution to amplification in the neuronal network explodes with the spectral radius R of the connectivity matrix $\boldsymbol{W}$. This is because the amplification of the first $N-1$ Schur units explodes with the variance ζ^2 of their feedforward interactions (Figure 2.3b) and that ζ^2 is directly related to R (Equation 2.9). Note that for $R > 1$ (to the right of the dashed vertical line in Figure 2.4a), the network of neurons is unstable. Although the concept of amplification in an unstable network is ill-defined, the "purely nonnormal" part of the total (infinite) amplification remains bounded. Indeed, the

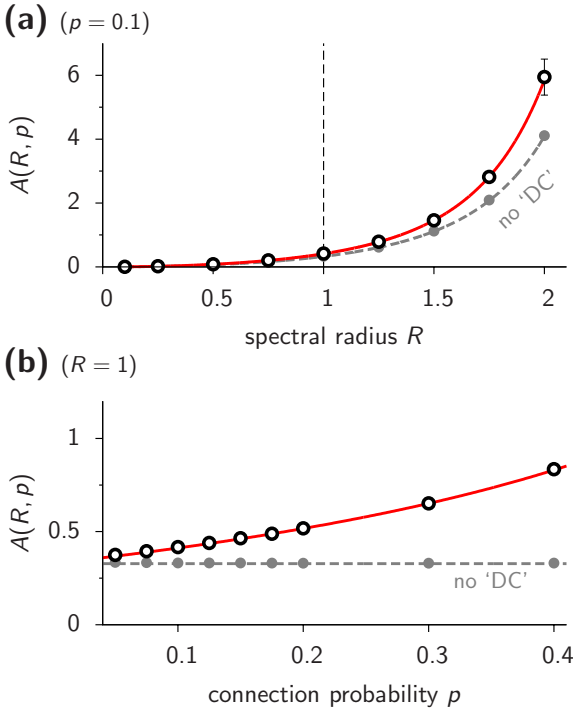


Figure 2.4: **Nonnormal amplification in random neuronal networks. (a)** The mean amount of purely nonnormal amplification $\langle A(\mathbf{T})\rangle \equiv A(R,p)$ is reported as a function of the spectral radius R of $\mathbf{W}$. Open circles denote the numerical solution of Equation 2.20 averaged over 20 randomly drawn connectivity matrices with connection density $p = 0.1$ and size $N = 500$. Errorbars denote the standard deviation over all trials. The red (upper) curve depicts the exact solution in Equation 2.17. The dashed grey (lower) curve and grey circles indicate the mean removal of Equation 2.11 applied to $\boldsymbol{W}$, which effectively removes the global macroscopic fluctuations of the entire population (labelled "no DC"). The dashed vertical line represents the limit of linear stability, beyond which the nonnormal part of amplification is still well-defined. **(b)** Same as in (a), now as a function of the connection density p for a fixed $R = 1$. In both (a) and (b), parameters p and R fully determined the value $\pm w_0/\sqrt{N}$ of the nonzero synaptic weights as $w_0 = R/\sqrt{p(1-p)}$ (cf. Equation 2.4).

purely feedforward network $\boldsymbol{T}$ derived from the Schur decomposition of $\boldsymbol{W}$ is itself always stable, since zero is the only eigenvalue of $\boldsymbol{T}$. The instability in $\boldsymbol{W}$ arises from purely normal effects, when the real part of one eigenvalue of $\boldsymbol{W}$ exceeds unity so that dynamical slowing becomes infinite.

Equation 2.16 confirms what we had previously discussed at the end of section 2.3: the last Schur unit has temporal fluctuations $\boldsymbol{v}\cdot\boldsymbol{x}(t)$ of variance $\mathcal{O}(N)$. Those fluctuations thus make up for a finite fraction of the total nonnormal amplification (the last term in Equation 2.17) as $N\to\infty$. Because the last Schur vector is the normalised uniform spatial pattern $(1,\ldots,1)/\sqrt{N}$, the variance of the overall population activity $\mu(t)\equiv\sum x_i(t)/N = \sqrt{N}(\boldsymbol{x}\cdot\boldsymbol{v}(t))$ is of order 1. As we had foreseen in section 2.3, one can restore the $1/N$ scaling of the these "DC" fluctuations $\langle\mu^2(t)\rangle$ by performing the operation of Equation 2.11 on the connectivity matrix $\boldsymbol{W}$, i.e. subtracting a common constant from all excitatory weights (including zero weights) to make sure that they average to zero, and adding the same constant to all inhibitory weights with the same purpose. This situation is depicted by the grey curves in Figure 2.4. Figure 2.4b shows that only these DC fluctuations depend on the connectivity density p.

Overall, Figure 2.4a allows us to draw two important conclusions. On the one hand, if the level of dynamical slowing is to be kept low ($R \ll 1$), only modest levels of amplification can be achieved (see the small amount of nonnormal amplification on the l.h.s. of the dashed vertical line). For example, if no mode is to decay with more than twice the single neuron time constant ($\mathrm{Re}(\lambda) < 1/2$), the average variance cannot exceed that of a disconnected network by more than 10%. On the other hand, the nonnormal contribution to amplification explodes with increasing R, i.e. with increasing synaptic strengths if the connection density is taken fixed. This suggests that strong transient amplification without dynamical slowing can only be achieved in structured, "less random" networks. The structure must allow the synaptic couplings to assume larger values without causing the eigenvalue spectrum of $\boldsymbol{W}$ to reach instability.

2.6 Different numbers of excitatory and inhibitory neurons

We now consider the biologically more plausible case of different numbers of excitatory and inhibitory neurons. Typical models of cortex assume fN excitatory neurons and $(1-f)N$ inhibitory neurons with $f = 0.8$ or similar. In this case, the eigenvalues λ are no longer uniformly scattered inside the disk of radius R in the complex plane[7], but become more concentrated in the middle following a radially symmetric density $\rho(|\lambda|)$ known analytically from (Rajan and Abbott, 2006) (Figure 2.5b, insets). As before, we consider the case where excitatory (resp. inhibitory) synaptic couplings are 0 with probability $(1-p)$, and $+w_E/\sqrt{N}$ (resp. $-w_I/\sqrt{N}$) otherwise. The global balance condition reads $fw_E = (1-f)w_I$. To impose a given spectral radius R, we set $w_E^2 = w_0^2(1-f)/f$ and $w_I^2 = w_0^2 f/(1-f)$ with $w_0^2 = R^2/p(1-p)$.

The results of section 2.3 regarding the variances in the Schur triangle have to be adjusted to accommodate these modifications. The derivation of ζ_0^2 is left unchanged, so that the couplings t_{Nj} onto the uniform mode $\boldsymbol{v}$ still have the variance given by Equation 2.8, which notably does not depend on f. Using Equation 2.5, we can then write down the empirical variance in the first $N-1$ rows of $\boldsymbol{T}$ as

$$\frac{2}{N(N-1)} \sum_{j<i<N} t_{ij}^2 = \frac{2}{N}\left(R^2 - \int_0^R r\rho(r)dr\right) \tag{2.18}$$

[7] Rajan and Abbott showed that this happens when the *variances* of the excitatory and inhibitory weights differ (the variances comprise both the zero and non-zero synapses). Decreasing the number of inhibitory neurons in a balanced network requires the strength of inhibition to be increased. In sparse networks like ours, this automatically makes the overall variance of the inhibitory synapses larger than that of excitatory synapses, hence the observed effect on the eigenspectrum.

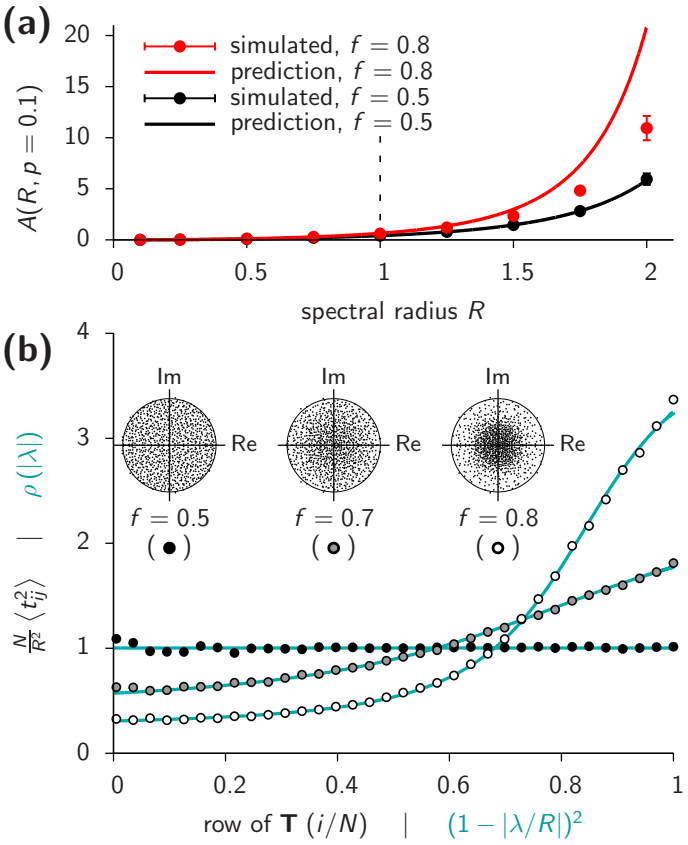


Figure 2.5: **Networks with different numbers of excitatory and inhibitory neurons. (a)** Nonnormal amplification as a function of the spectral radius R of $\boldsymbol{W}$, in sparse random balanced networks with fN excitatory and $(1-f)N$ inhibitory neurons, for $f = 0.5$ (black, lower) and $f = 0.8$ (red, upper). The connection density p was set to 0.1. The dashed vertical line represents the limit of linear stability, beyond which the nonnormal part of amplification is still well-defined. Solid circles were obtained by averaging the numerical solution of Equation 2.20 for 20 random matrices of size $N = 500$. Errorbars denote standard deviation over all trials. **(b)** Filled circles show the scaled variance $N\langle t_{ij}^2 \rangle / R^2$ of the non-zero Schur couplings in row i as a function of i/N and for three different values of f. These variances were computed by Schur-transforming 100 matrices of size $N = 200$, with $R = 1$ and $p = 0.1$. Cyan lines denote the density ρ of eigenvalues λ inside the unit disk (Rajan and Abbott, 2006), as a function of $(1 - |\lambda/R|)^2$. Insets show the eigenvalue spectra of three example matrices of size $N = 1000$.

Unfortunately, the *ensemble* variance $\left\langle t_{ij}^2 \right\rangle$ for fixed i and j is in general different from the average across matrix elements given by Equation 2.18. Indeed, contrary to the case $f = 0.5$ considered in section 2.3, the non-zero elements of $\boldsymbol{T}$ *no longer* have the same ensemble variance. Instead, $\left\langle t_{ij}^2 \right\rangle$ grows with row index $i < N$, and this profile interestingly matches the density of eigenvalues ρ [8], according to

$$\frac{N}{R^2}\left\langle t_{ij}^2 \right\rangle \quad = \quad \rho\left[R\left(1 - \sqrt{\frac{i}{N}}\right)\right] \quad \text{for} \quad j < i < N \tag{2.19}$$

This is depicted in Figure 2.5b.

In a feedforward network like that of Schur units considered here, a good strategy to generate greater amplification would be to give comparatively more power to the couplings onto

[8] This happens provided the eigenvectors are sorted in decreasing order of their corresponding eigenvalue moduli, prior to going through the Gram-Schmidt orthonormalisation process. This results in a unique Schur basis.

earlier nodes. This is because amplification builds up superlinearly along the feedforward chain (Figure 2.3), so that boosting early nodes exacerbates the avalanche effect (see also (Ganguli et al., 2008b)). Setting f to more than 0.5 does precisely the contrary: couplings onto early nodes become comparatively smaller in magnitude, as shown by the filled circles in Figure 2.5b. Therefore, simply replacing α^2/N in Equation 2.15 by the empirical variance of Equation 2.18 yields an over-estimation of the true amplification in the first $N-1$ Schur units (compare the red line with the red circles in Figure 2.5a). We found it difficult to incorporate this variance profile $\left\langle t_{ij}^2 \right\rangle$ into the derivation of appendix section 2.A, so we can only consider as accurate the results of numerical simulations.

The conclusions reached at the end of section 2.5 do not change significantly under the more realistic assumption of $f = 0.8$. Although amplification almost doubles relative to $f = 0.5$, it remains very weak in the stable regime (to the left of the dashed vertical line in Figure 2.5a), confirming that amplification can only come with substantial dynamical slowing when connections are drawn at random.

2.7 Example of network structure for nonnormal amplification

Here we show that random networks can be minimally structured in such a way that strong nonnormal amplification occurs already in the fast dynamical regime. We exploit the fact that correlations in the connectivity matrix can modify the shape of the eigenvalue spectrum. Symmetrising (or anti-symmetrising) $\boldsymbol{W}$ has been shown to generate elliptical (as opposed to circular) eigenspectra, in the case of "centered" matrices where the distinction between excitatory and inhibitory neurons is not made (Sommers et al., 1988). Here we consider a modification of the sparse neural matrices studied in section 2.3 that achieves this slimming effect in the case of balanced networks (see the insets in Figure 2.6a). All non-zero entries assume a value $\pm w_0/\sqrt{N}$, the sign depending on the excitatory versus inhibitory nature of the presynaptic neuron. Whether a connection exists (non-zero entry) is decided as follows. Connection w_{ij} with $i \geq j$ exists with probability p. If $i \neq j$, the reciprocal connection w_{ji} then exists with probability $p + c_{ij}(1-p)$ if w_{ij} exists too, or with probability $p(1-c_{ij})$ if it does not. In comparison to the random networks considered above (Equation 2.3), this connectivity scheme preserves the mean weight $\bar{w} \equiv \left\langle w_{ij} \right\rangle = \pm p w_0/\sqrt{N}$ as well as the weight variance $\left\langle (w_{ij} - \bar{w})^2 \right\rangle = p(1-p)w_0^2/N$ while giving full control over their normalized covariance c_{ij}. Note that c_{ij} can assume positive values as high as $c_{\max} = 1$, in which case all connections are bidirectional. However, c cannot go below $c_{\min} = -p/(1-p)$, which stems from the sparsity condition that imposes a certain degree of symmetry in $\boldsymbol{W}$: because both w_{ij} and w_{ji} are zero with high probability, they will often be null together,

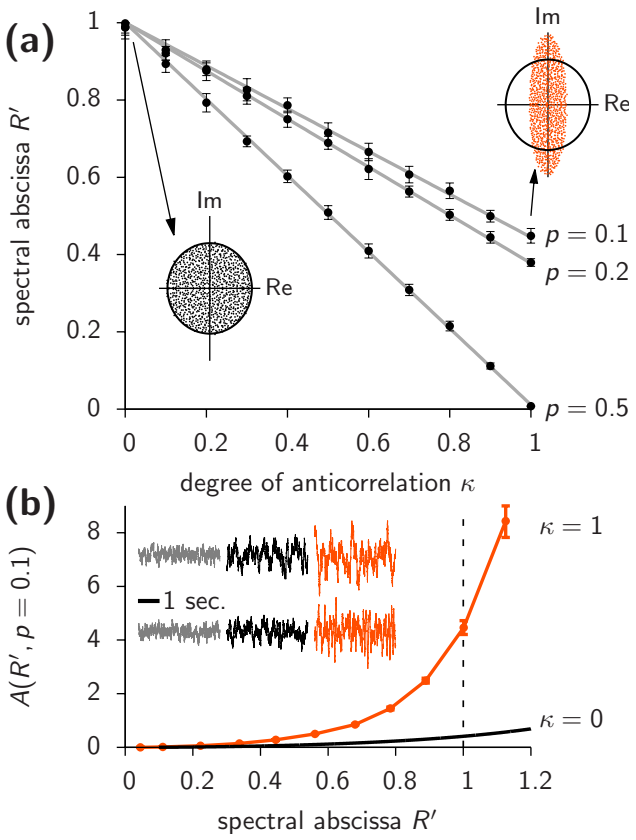


Figure 2.6: **Example of a network structure that favors nonnormal amplification: unidirectional vs. bidirectional synaptic connections.** **(a)** We varied the degree of anticorrelation between reciprocal weights in the connectivity matrix, as the fraction κ of the maximum value it can assume, which is dictated by the connection sparsity (see text). This caused the eigenspectrum to stretch more and more along the vertical axis (compare the two insets), effectively decreasing the spectral abscissa R' (black filled circles). Empirical data was obtained from numerically computing the eigenvalues of 20 different matrices of size $N = 500$. Errorbars denote standard deviations over all trials. Gray lines are linear fits. **(b)** Nonnormal amplification as a function of the spectral abscissa R'. When all connections between an excitatory (E) and an inhibitory (I) cell are made reciprocal, while all E $\to$ E and I $\to$ I connections are kept unidirectional (orange (upper) curve, corresponding to $\kappa = 1$ in (a)), stronger amplification is obtained in the fast dynamical regime ($R' \ll 1$). The black (lower) curve is here reproduced from Figure 2.4 (purely random case, $\kappa = 0$) for comparison. The inset displays examples of 4-second snapshots of activity in a disconnected network (left), a random network (middle, $\kappa = 0$), and a maximally (though not fully) antisymmetric network (right, $\kappa = 1$). The spectral abscissa was set to $R' = 0.9$. Traces were obtained from a direct simulation of Equation 2.1, and are shown here only for two randomly chosen neurons.

meaning that they cannot be fully anti-correlated. The limit case $c = c_{\min}$ corresponds to the complete absence of reciprocal connections. Since we aim at tilting $\boldsymbol{W}$ towards antisymmetry, we choose $c_{ij} = \kappa c_{\min}$ when neurons i and j are of the same type, and $c_{ij} = \kappa c_{\max}$ when the two neurons have different types. Thus $0 < \kappa < 1$ parameterises the degree of antisymmetry in $\boldsymbol{W}$. As can be seen in Figure 2.6a, increasing κ effectively decreases the spectral abscissa $R' = \max_\lambda \mathrm{Re}(\lambda)$, although it is designed not to affect the overall connectivity "power" $\sum w_{ij}^2$ which *is* the relevant quantity for amplification. Thus, for a fixed level of dynamical slowing (i.e. fixed R'), antisymmetric connectivity matrices can assume larger weight strengths and thereby yield stronger nonnormal amplification than their random counterparts, as depicted in Figure 2.6b. Finally, note that a matrix with $\kappa = 1$ is *not* purely antisymmetric ($\boldsymbol{W}^\dagger \neq -\boldsymbol{W}$). In fact, neural connectivity matrices can never be fully antisymmetric, because of the constraint that neurons can be only excitatory or only

inhibitory. This is an advantageous restriction here, because a fully antisymmetric matrix – just like a fully symmetric one – is in fact a normal operator that cannot support transient amplification.

2.8 Discussion

The nonnormal nature of the neuronal connectivity could play a major role in the functional dynamics of cortical networks. It can allow fast transients to develop along well-defined activity motifs stored in the pattern of synaptic efficacies. In networks with locally dense connectivity, the balance between excitation and inhibition has been shown to generate amplification of this type, accordingly termed "balanced amplification" (Murphy and Miller, 2009). We have revisited this feature in sparse balanced networks in which any two neurons are connected randomly with some probability. Random networks had already been studied in terms of their pseudospectrum (Trefethen and Embree, 2005), which only provides bounds on amplification. We have chosen a more direct approach and assessed nonnormality in terms of its functional impact in networks driven by stochastic external input. We have explicitly calculated the strength of the activity fluctuations that can only be attributed to the nonnormality of the recurrent connectivity. We found nonnormal amplification to be very weak, concluding that the only way to obtain large amplification in random networks is to allow for significant dynamical slowing. If the dynamics are to be kept fast, then the connectivity needs some structuring, so as to allow synaptic weights to take up larger values and to discourage the emergence of large positive eigenvalues. We have given an example of minimal network structure, namely connection antisymmetry, that achieves precisely this. More adaptive ways of shaping the connectivity, such as synaptic plasticity, could also be considered. In particular, inhibitory synaptic plasticity has recently been shown to suppress the attractor dynamics of a few activity motifs embedded in a spiking network, while still permitting their transient recall (Vogels et al., 2011).

Nonnormal amplification could provide a mechanistic account for the often reported transient nature of both spontaneous and evoked activity in primary sensory cortices. Moreover, from a functional viewpoint, amplification without slowing could be a highly relevant feature in areas involved in the processing of fast-changing signals. If transient amplification by the synaptic connectivity is meant to allow past experience to be reflected in the responses to sensory stimuli (see e.g. (Fiser et al., 2010)), then it is quite reassuring that random networks are poor amplifiers, for it implies that nothing can be amplified that has not been learned.

Here we have focused on spontaneous activity, i.e. on the fluctuations elicited by isotropic external noise that is totally uninformed of the frozen structure of the connectivity matrix.

The equivalent triangular form of a nonnormal connectivity matrix suggests that neuronal networks should be more sensitive along some input directions than along others, so they could still respond vigorously though transiently to some carefully chosen input patterns (evoked activity). The first Schur mode, for example, is indeed such a preferred pattern (Ganguli et al., 2008b). This anisotropy prompts two important questions. First, how many different (orthogonal?) directions of high sensitivity does a network possess? Similarly, in how many distinguishable directions can the network amplify those preferred input signals? These quantities taken together could define the "nonnormal information capacity" of a network, reminiscent of the concept of memory capacity in attractor networks.

We have assumed here a simple network topology of the Erdős-Rényi type, whereas brain networks are often more heterogeneous (Sporns, 2011), e.g. small-world and/or scale-free (Shefi et al., 2002; Eguíluz et al., 2005). The graph topology is known to affect dynamical properties such as correlations and network synchronisation (Roxin, 2011) or performance in attractor tasks (de Franciscis et al., 2011). The width of the out-degree distribution could prove particularly important to the phenomenon we study here, since it modulates the amount of shared input between cells, and therefore also the magnitude of pairwise correlations (Pernice et al., 2011) that can in turn source amplification. Although more complicated topologies fall outside the scope of our study, it would be interesting to see how they affect the nonnormal contribution to amplification, as opposed to how they dictate the eigenvalue spectrum of the adjacency matrix (see e.g. (Goh et al., 2001; Grabow et al., 2012) for spectral analyses).

Finally, our analysis has revealed that the nonnormality of balanced networks is to a large extent reflected in large "DC" fluctuations. This seems to be a general feature of networks in which neurons can either be excitatory or inhibitory, but not of a mixed type (Kriener et al., 2008). It is somewhat disappointing that however strong activity fluctuations are in individual neurons, they always comprise a finite fraction of common variability. This is because the variance of the overall population activity is of the same order as the activity variance of the individual neurons (Equation 2.16). Should computations exploit the fluctuations along the remaining $N-1$ degrees of freedom of the network, complications in decoding the current network state would most certainly arise from a single dimension dominating the dynamics. However, we wish to point out that these large DC fluctuations are in fact a direct consequence of the *exact* excitation-inhibition balance considered here. We show in appendix section 2.C that when inhibition *dominates* over excitation, the variance of the population activity becomes suddenly inversely proportional to the network size. Furthermore, the mean pairwise correlation coefficient in the network scales similarly, and thus vanishes in large networks unless the E-I balance is exact. Note that this phenomenon is *not* mediated by a destruction of the strong feedforward link from the global balance disruption $\boldsymbol{d}$ onto the DC

mode $\boldsymbol{v}$, as described at the end of section 2.3. Increasing the overall amount of inhibition does preserve this strong link, but cancels its amplifying effect by imposing an equally strong negative feedback from the DC mode onto itself (see appendix section 2.C). This dynamic cancellation of fluctuations and correlations was already shown to arise in balanced networks of spiking neurons (Renart et al., 2010). Our results obtained for linear networks therefore suggest it may be a very general feature of inhibition-dominated balanced networks, and that fine-tuning the balance until it becomes exact (Vogels et al., 2011) may strongly affect the dynamics of the network and the resulting correlation structure.

2.A Amplification in random triangular networks

In this appendix we derive an exact expression for amplification in random strictly triangular networks with linear stochastic dynamics as in Equation 2.1, where the non-zero elements of the coupling matrix $\boldsymbol{T}$ are drawn from an arbitrary distribution with zero mean and variance α^2/N where N is the network size. Though no closed-form solution is known for the zero time lag covariance matrix $\boldsymbol{\Sigma}$, we know from the theory of multidimensional Ornstein-Uhlenbeck processes that it satisfies the so-called Lyapunov equation (Gardiner, 1985)

$$(\boldsymbol{T} - \mathbb{1})\,\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\left(\boldsymbol{T}^{\dagger} - \mathbb{1}\right) = -\tau\sigma_{\xi}^{2}\mathbb{1} \tag{2.20}$$

Equating component $(i, j < i)$ on both sides of Equation 2.20 yields:

$$\sigma_{ij} = \frac{1}{2}\sum_{k=1}^{i-1} t_{ik}\sigma_{jk} + \frac{1}{2}\sum_{k=1}^{j-1} t_{jk}\sigma_{ik} \tag{2.21}$$

and equating the diagonal term (i, i) on both sides gives the variance of Schur mode i:

$$\sigma_{ii} = \frac{\tau\sigma_{\xi}^{2}}{2} + \sum_{j=1}^{i-1} t_{ij}\sigma_{ij} \tag{2.22}$$

Combining Equation 2.21 and Equation 2.22 yields

$$\sigma_{ii} = \frac{\tau\sigma_{\xi}^{2}}{2} + \frac{1}{2}\sum_{j=1}^{i-1} t_{ij}\left(\sum_{k=1}^{i-1} t_{ik}\sigma_{jk} + \sum_{k=1}^{j-1} t_{jk}\sigma_{ik}\right) \tag{2.23}$$

in which σ_{jk} and σ_{ik} are to be recursively obtained from Equation 2.21 with proper replacement of indices. We would like to calculate the expected value over the t_{ij} coefficients, i.e. over multiple realisations of random matrix $\boldsymbol{T}$. Explicitly expanding the sums will reveal cross-terms like $\langle t_{ij}t_{k\ell}\rangle$. Those vanish if $i \neq k$ or $j \neq \ell$, because the coupling coefficients

are taken to be uncorrelated. The only remaining terms will be powers of the variance α^2/N. Here we seek a truncation to order α^4/N^2. Let us calculate:

$$\begin{aligned}\langle\sigma_{ii}\rangle = \frac{\tau\sigma_\xi^2}{2} + \frac{1}{2}\sum_{j=1}^{i-1}\sum_{k=1}^{i-1}\left\langle t_{ij}t_{ik}\sigma_{jk}\right\rangle \\ +\frac{1}{2}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\left\langle t_{ij}t_{jk}\sigma_{ik}\right\rangle\end{aligned} \tag{2.24}$$

Because the network of Schur modes is purely feedforward, the cross-covariance σ_{jk} for $(j,k) < i$ is independent of the coupling coefficients t_{ij} and t_{ik}, thus $\left\langle t_{ij}t_{ik}\sigma_{jk}\right\rangle = \left\langle t_{ij}t_{ik}\right\rangle\left\langle\sigma_{jk}\right\rangle$. The only non-vanishing term in the first double-sum is therefore obtained for $k = j$, giving

$$\langle\sigma_{ii}\rangle = \frac{\tau\sigma_\xi^2}{2} + \frac{\alpha^2}{2N}\sum_{j=1}^{i-1}\langle\sigma_{jj}\rangle + \frac{1}{2}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\left\langle t_{ij}t_{jk}\sigma_{ik}\right\rangle \tag{2.25}$$

Let us expand the expression in the second double-sum using Equation 2.21:

$$\begin{aligned}\left\langle t_{ij}t_{jk}\sigma_{ik}\right\rangle = \frac{1}{2}\sum_{\ell=1}^{i-1}\left\langle t_{ij}t_{jk}t_{i\ell}\sigma_{k\ell}\right\rangle \\ +\frac{1}{2}\sum_{\ell=1}^{k-1}\left\langle t_{ij}t_{jk}t_{k\ell}\sigma_{i\ell}\right\rangle\end{aligned} \tag{2.26}$$

As above, the first sum vanishes except for $\ell = j$. Should one continue and expand the second sum, one would receive terms of order α^6/N^3 and more which are discarded here (see above). Hence

$$\left\langle t_{ij}t_{jk}\sigma_{ik}\right\rangle = \frac{\alpha^2}{2N}\left\langle t_{jk}\sigma_{jk}\right\rangle + \cdots \tag{2.27}$$

Using similar arguments, we expand $\left\langle t_{jk}\sigma_{jk}\right\rangle$ to order α^2/N and receive:

$$\left\langle t_{jk}\sigma_{jk}\right\rangle = \frac{\alpha^2}{2N}\left\langle\sigma_{kk}\right\rangle + \cdots \tag{2.28}$$

From Equation 2.25 it therefore follows that

$$\begin{aligned}\langle\sigma_{ii}\rangle = \frac{\tau\sigma_\xi^2}{2} + \frac{\alpha^2}{2N}\sum_{j=1}^{i-1}\langle\sigma_{jj}\rangle \\ + \frac{\alpha^4}{8N^2}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\langle\sigma_{kk}\rangle\end{aligned} \tag{2.29}$$

Defining $f_i = 2\langle\sigma_{ii}\rangle/(\sigma_\xi^2\tau)$, we end up with a recursive equation for the build-up of relative variance down the feedforward network of Schur modes:

$$f_i = 1 + \frac{\alpha^2}{2N}\sum_{j=1}^{i-1} f_j + \frac{\alpha^4}{8N^2}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} f_k \tag{2.30}$$

Now we define $x = i/N$ (thus $0 \leq x \leq 1$) and rewrite Equation 2.30 as

$$\begin{aligned} f_{xN} = 1 &+ \frac{\alpha^2 x}{2i} \sum_{j=1}^{i-1} g\left(\frac{xj}{i}\right) \\ &+ \frac{\alpha^4 x^2}{8i^2} \sum_{j=1}^{i-1} \sum_{k=1}^{j-1} g\left(\frac{xk}{i}\right) \end{aligned} \tag{2.31}$$

In the limit $N \to \infty$ with constant $x = i/N$ ratio, the sums on the r.h.s. converge to their corresponding Riemann integrals, endowing f_{xN} with a proper limit $g(x)$:

$$\begin{aligned} g(x) = 1 &+ \frac{\alpha^2 x}{2} \int_0^1 g(xs)\, ds \\ &+ \frac{\alpha^4 x^2}{8} \int_0^1 ds \int_0^1 ds' \Theta(s - s')\, g(xs') \end{aligned} \tag{2.32}$$

where Θ is the Heaviside function. This convergence stems from the $1/N$ scaling of the variance α^2/N. Using straightforward changes of variables ($s \mapsto s/x$), we end up with an integral equation for g, the continuous variance profile along the (now infinitely large) network of Schur patterns:

$$g(x) = 1 + \frac{\alpha^2}{2} \int_0^x g(s)\, ds + \frac{\alpha^4}{8} \int_0^x ds \int_0^s ds' g(s') \tag{2.33}$$

Differentiating Equation 2.33 twice with respect to x yields a second-order differential equation for g

$$g''(x) = \frac{\alpha^2}{2} g'(x) + \frac{\alpha^4}{8} g(x) \tag{2.34}$$

with initial conditions $g(0) = 1$, $g'(0) = \alpha^2/2$, and $g''(0) = 3\alpha^4/8$. The solution is precisely $g^{\mathrm{LB}}(x)$ given in Equation 2.12 of the main text. It is only a lower-bound on the true variance profile $g(x)$ since all the higher-order terms in α^2 that we have neglected are positive. This approximation proves reasonable for $\alpha^2 < 3$ as shown in Figure 2.3a (dashed blue lines). Further integrating over x yields a lower-bound on nonnormal amplification $A_0(\alpha^2) \equiv \int_0^1 g(x)\, dx - 1$ (Figure 2.3b, dashed blue line):

$$\begin{aligned} A_0^{\mathrm{LB}}(\alpha^2) = &\frac{2}{\alpha^2\sqrt{3}} \exp\left(-\frac{(\sqrt{3}-1)\alpha^2}{4}\right) \\ &\times \left[\exp\left(\frac{\sqrt{3}\alpha^2}{2}\right) - 1\right] - 1 \end{aligned} \tag{2.35}$$

Instead of truncating $\langle \sigma_{ii} \rangle$ to order α^4, one can also decide to start again from Equation 2.24 and keep all terms up to order n. This requires careful counting, and results in a differential equation of order n, reading

$$g^{(n)}(x) = \frac{\alpha^2}{2} \sum_{k=0}^{n} C_k \left(\frac{\alpha^2}{4}\right)^k g^{(n-k-1)}(x) \tag{2.36}$$

where $C_k = (2k)!/[k!(k+1)!]$ is the k^{th} Catalan number. Assuming $g(x)$ can be written for $0 \leq x \leq 1$ as a convergent power series

$$g(x) = \lim_{K\to\infty} \sum_{k=0}^{K} \beta_k x^k \tag{2.37}$$

and equating $g^{(k)}(0)$ in both Equation 2.36 and Equation 2.37 yields the results of Equation 2.13 – Equation 2.15.

2.B Variance of the DC component

The last Schur mode is fed by the activities of all previous Schur vectors, weighted by couplings with variance ζ_0^2/N. The same calculation that led to Equation 2.30 in this case leads to

$$f_N = 1 + \frac{\zeta_0^2}{2} \sum_{j=1}^{N-1} f_j + \frac{\zeta_0^2 R^2}{8N} \sum_{j=1}^{N-1} \sum_{k=1}^{j-1} f_k + \cdots \tag{2.38}$$

which can be rewritten as

$$\frac{f_N}{N} = \frac{1}{N} + \frac{\zeta_0^2}{R^2} \left(\frac{R^2}{2N} \sum_{j=1}^{N-1} f_j + \frac{R^4}{8N^2} \sum_{j=1}^{N-1} \sum_{k=1}^{j-1} f_k + \cdots \right) \tag{2.39}$$

where the sums were previously calculated in the limit $N \to \infty$ (Equation 2.30 – Equation 2.37). We thus recover

$$\lim_{N\to\infty} \frac{f_N}{N} = \frac{\zeta_0^2}{R^2} \left[g\left(1\right) - 1 \right] \tag{2.40}$$

With ζ_0^2 given by Equation 2.8 we arrive at Equation 2.16 of the main text.

2.C Exactly balanced vs. inhibition-dominated networks

In this paper, we have considered connectivities in which weights were either zero or $\pm w_0/\sqrt{N}$, the $\pm$ sign depending on the excitatory vs. inhibitory nature of the presynaptic neuron (Equation 2.3). Furthermore, the number of cells of both types was identical. The total inhibitory synaptic strength thus exactly matched its excitatory counterpart. In this appendix, we wish to show that if the non-zero inhibitory weights are stronger, i.e. $-\gamma w_0/\sqrt{N}$ with $\gamma > 1$, the dynamics of the overall population activity is strongly affected.

We have seen that the "DC" mode $\boldsymbol{v} = (1, \ldots, 1)/\sqrt{N}$ is an eigenvector of $\boldsymbol{W}$. Let λ_v denote the associated eigenvalue, which quantifies the effective decay rate of the DC component in the network of neurons. If the E-I balance is exact ($\gamma = 1$) as assumed throughout the paper, then $\lambda_v = 0$. More generally, however, one can calculate

$$\lambda_v \quad = \quad -\frac{pw_0(\gamma - 1)}{2} \cdot \sqrt{N} \tag{2.41}$$

We see there is an unexpected scaling that the exact balance was hiding : $-\lambda_v \sim \mathcal{O}(\sqrt{N})$. Note that all other eigenvalues are now scattered inside the disk of radius

$$R \quad = \quad w_0\sqrt{\frac{(1+\gamma^2)p(1-p)}{2}} \tag{2.42}$$

though no longer uniformly so since the variance of the inhibitory and excitatory weights now differ by a factor of γ^2 (Rajan and Abbott, 2006). Having kept the focus of this paper on nonnormal effects, we have intentionally set aside the contributions of the eigenvalues to the overall amplification in the network. When $\lambda_v = 0$ (perfect balance), our prediction that the average population activity $\mu(t) \equiv \sum x_i(t)/N$ should have a variance of order $\mathcal{O}(1)$ was justified : the last Schur unit corresponding to this DC indeed receives $N-1$ contributions of order $\mathcal{O}(1)$, and its decay time constant is simply $\tau \sim \mathcal{O}(1)$, yielding $\mathrm{var}[\mu(t)] \sim \mathcal{O}(1)$. When inhibition dominates ($\gamma > 1$), the DC component suppresses itself via a negative feedback that scales with $\sqrt{N}$, yielding a very short decay time constant $\tau/(1-\lambda_v) \sim \mathcal{O}(1/\sqrt{N})$ whose deviation from τ can no longer be neglected. To see what the implications of this scaling are for the variance of $\mu(t)$, let us reduce the dynamics of the DC to the following set of N stochastic differential equations:

$$\begin{aligned} dy_i &= -\frac{dt}{\tau}y_i + \sqrt{\frac{2}{\tau}}\,d\xi_i \quad \text{for} \quad 1 \le i < N \\ dy_N &= \frac{dt}{\tau}\left(-(1-\lambda_v)y_N + \sum_{i=1}^{N-1}\varepsilon_i x_i\right) + \sqrt{\frac{2}{\tau}}d\xi_N \end{aligned} \tag{2.43}$$

Here $y_1, \ldots, y_{N-1}$ model the first $N-1$ Schur units independently, with the appropriate noise terms such that they achieve a variance of one (corresponding to the limit of small amplification). They feed y_N – which models the activity of the last Schur unit, i.e. the DC component $\mu(t)\sqrt{N}$ – with couplings ε_i such that $\sum \varepsilon_i^2/N = \zeta_0^2$. We calculate the coupling variance ζ_0^2 the same way we did in section 2.3:

$$\zeta_0^2 \quad = \quad \frac{p^2 w_0^2(1+\gamma^2)}{2} \tag{2.44}$$

The variance $\mathrm{var}[\mu(t)]$ of the overall neuronal population activity, here modeled by $\mu(t) \approx y_N(t)/\sqrt{N}$, is given by standard Ornstein-Uhlenbeck theory:

$$\mathrm{var}(\mu(t)) \quad = \quad \frac{1}{N(1-\lambda_v)}\left[1 + \frac{N\zeta_0^2}{2-\lambda_v}\right] \tag{2.45}$$

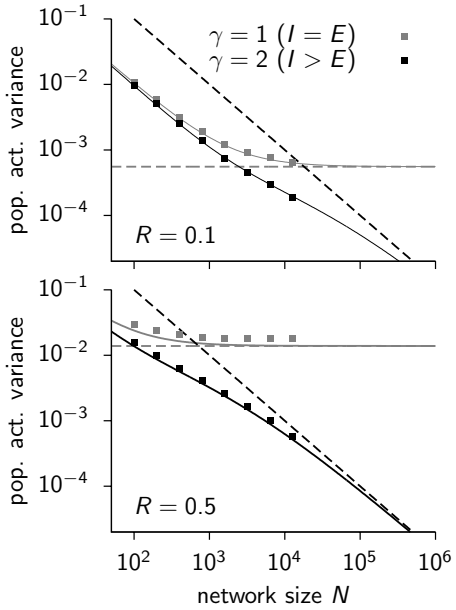


Figure 2.7: **Temporal fluctuations of the overall population firing rate in a balanced neuronal network**. The variance of the average population activity $\mu(t) = \sum x_i(t)/N$ is reported as a function of the network size N, in logarithmic scale. When inhibition perfectly balances excitation ($\gamma = 1$), the variance is asymptotically independent of the network size (gray). When inhibition dominates ($\gamma > 1$), it scales with $1/N$ (black). The solid lines denote the approximation in Equation 2.45. The dashed lines indicate the asymptotics (Equation 2.46). Points denote the empirical variance obtained by simulating Equation 2.1 for 100 seconds, for a neuronal network constructed as specified in section 2.3 with connectivity density $p = 0.1$. The spectral radius was set to $R = 0.1$ (top plot) and $R = 0.5$ (bottom plot).

Although we have neglected amplification and correlations in the first $N - 1$ Schur units, Equation 2.45 does provides a good intuition for how the mean population activity $\mu(t) = y_N(t)/\sqrt{N}$ scales with the network size N, and provides a good qualitative match to numerical results even for a non-negligible spectral radius $R = 0.5$ (Figure 2.7).

The asymptotics of $\mathrm{var}[\mu(t)]$ are given by

$$\mathrm{var}[\mu(t)] \quad \sim \quad \begin{cases} \dfrac{p^2 w_0^2}{2} & \text{if} \quad \gamma = 1 \\ \dfrac{2(1+\gamma^2)}{N(\gamma-1)^2} & \text{if} \quad \gamma > 1 \end{cases} \tag{2.46}$$

Thus, when inhibition dominates over excitation ($\gamma > 1$), the fluctuations of the overall population activity vanish for large networks, which was already shown in (Renart et al., 2010) for inhibition-dominated networks of spiking neurons. In contrast, fine tuning the connectivity such that the balance becomes exact ($\gamma = 1$) opens the possibility for these fluctuations to subsist in arbitrarily large networks. This has profound consequences for the mean pairwise correlation $\bar{r} \equiv \sum_{i \neq j} \mathrm{cov}[x_i(t), x_j(t)]/N^2$, as seen from the following identity

$$\bar{r} \quad = \quad \mathrm{var}[\mu(t)] - \frac{1}{N^2} \sum_i \mathrm{var}[x_i(t)] \tag{2.47}$$

We have seen that the average variance $\mathrm{var}[x_i(t)]$ in the individual neurons (i.e. amplification as we define it) is $\mathcal{O}(1)$. Thus, Equation 2.47 implies that $\bar{r}$ scales with N in the same way $\mathrm{var}[\mu(t)]$ does: either $\mathcal{O}(1)$ if the balance is perfect, or $\mathcal{O}(1/N)$ if inhibition dominates.

CHAPTER 3

Amplification and rotational dynamics in inhibition-stabilized cortical circuits

The technical findings of chapter 2 can be summarized as follows. When all the connections of a random balanced network are strenghtened, the network becomes unstable *before* the effect of the synaptic strengthening is felt through nonnormal (transient) amplifying effects (c.f. Figure 6.4). However, one can abstract away the instability and calculate the amount of amplification that would be contributed by the hidden feedforward connectivity if one could, by some magic, anihilate the large unstable eigenvalues of the connectivity matrix and make the network stable.

In this chapter, we show this is (almost) possible. Possible, because unstable balanced random networks with large synaptic weights can indeed be stabilized by properly tuned inhibition. "Almost", because in doing so, some degree of nonnormality is lost, and the amplification that survives the stabilization procedure is weaker than expected from our calculations of chapter 2.

The chapter is organized as a journal article, although it is not yet submitted (but will be soon). The paper's core theoretical component is a procedure for **optimal fine-tuning of the inhibitory synapses** of an unstable network, with the aim of **stabilizing** the recurrent dynamics. The procedure takes the form of an iterative update rule for the inhibitory synapses, which is directly inspired from recent advances in control theory. The technical details of the update rules are postponed to the Supplemental Data, where we also show how it is

approximated by a class of local plasticity rules.

From the optimization procedure we derive a series of results of relevance to neuroscience and our understanding of the dynamical regime in which the cortex operates. The main question is: **can the behaviour of inhibition-stabilized networks be related to known aspects of the cortical neurophysiology?**

3.1 Introduction

The generation of motor patterns has been the focus of several recent experimental studies (Churchland et al., 2010a, 2012). In a typical experiment, sketched in Figure 3.1A, a monkey is asked to prepare a particular arm movement, but not to initiate it before a go cue is delivered. Recordings have shown that, during the delay period, motor and pre-motor cortical areas transition from spontaneous firing activity into a movement-specific "preparatory" state, in which they remain until the go cue is issued. According to a recent proposal (Shenoy et al., 2011), motor populations could act as generic dynamical systems that different initial states would drive into different patterns of collective dynamics. In this view, planning a movement would require making sure the system arrives at the right initial condition by the time the movement must be triggered. When released, the population dynamics would then elicit the correct movement.

The above view, however, does not make any specific claim regarding the type of dynamical system, or neuronal network, suitable for movement generation. To achieve complex movements, intuitively, one may want the system to produce complex dynamics. In fact, single-neuron dynamics following the go cue are indeed both spatially and temporally complex, with multiphasic single-cell firing rate responses resembling the toy traces of Figure 3.1B (Churchland and Shenoy, 2007). Population transients last for only a few hundred milliseconds, and are characterized by large deviations from spontaneous firing rates on the single-cell level (Churchland et al., 2012). This type of dynamics is intriguing: the system is apparently highly excitable from the initial condition set up by the preparatory period, while also being stable and able to return to rest after a short while. How cortical networks could generate complex transient amplification of this sort through recurrent interactions is still poorly understood.

Here we address the mechanistic underpinnings of such transient collective behavior in rate models of balanced cortical dynamics, with an emphasis on network connectivity. Randomly connected balanced networks, though having complex connectivity, do not comply with the requirements set by the data. Indeed, weakly coupled random networks cannot produce

the substantial transient departure from background activity observed in the experiments (Hennequin et al., 2012). Strongly coupled random networks with their inherent chaotic dynamics, on the other hand, display complex behavior but do not capture the transient nature of movement-related activity (Sompolinsky et al., 1988; Rajan et al., 2010). Moreover, in the scenario underlined above where the initial condition is supposed to dictate the subsequent evolution of the system, chaotic behavior with high sensitivity to initial perturbations would seem ill-suited. Ideally, one would need strong and complex recurrent connectivity to coexist with stable dynamics. Sussillo and Abbott (2009) came up with an elegant solution: chaos can be controlled through the introduction of an appropriate feedback loop, allowing for the generation of stable trajectories. While we do not rule out such an attractive possibility, we explore here an alternative mechanism which also exploits a feedback loop but produces a different type of dynamics. We hypothesize that motor cortical circuits have strong and complex excitatory recurrent connectivity, but are stabilized by adequate recurrent inhibition. We call this new class of balanced networks "cISNs", or complex Inhibition-Stabilized Networks, in analogy with the concept of ISN introduced by Tsodyks et al. (1997) and recently developed by Ozeki et al. (2009). The goal of our study is *not* to study how the brain may learn appropriate inhibitory feedback, e.g. through inhibitory synaptic plasticity (Vogels et al., 2011; Luz and Shamir, 2012); instead, we provide a principled way of engineering cISNs with plausible connectivity and focus our study on their dynamical behavior.

As it turns out, cISNs transiently amplify a rich array of network states. The network activity can be forced to arrive at one of those states by the end of the preparatory period through the delivery of an appropriate external input. Upon a go cue, the input is withdrawn and the network is left to evolve freely, eliciting transient single-neuron and collective dynamics that match the data well. In particular, we reproduce the recently uncovered phenomenon of rotational ensemble dynamics following the go cue (Churchland et al., 2012). Additionally, muscle activities may be read out from these noisy transients to yield complex movements.

Interestingly, cISNs connect several previously disparate aspects of balanced cortical dynamics. The mechanism that underlies the generation of large transients here is a more general form of "Balanced Amplification" (Murphy and Miller, 2009), which was previously discovered in the context of visual cortical dynamics. Furthermore, during spontaneous activity in inhibition-stabilized networks, a detailed balance of excitatory and inhibitory inputs to single cells is established that is much finer than expected from shared population fluctuations (Vogels and Abbott, 2009; Okun and Lampl, 2008; Cafaro and Rieke, 2010). Overall, our results demonstrate the possibility for balanced cortical circuits to elicit transients of large amplitude along many different directions in state space, thus going beyond the transmission of information through population-averaged firing rates.

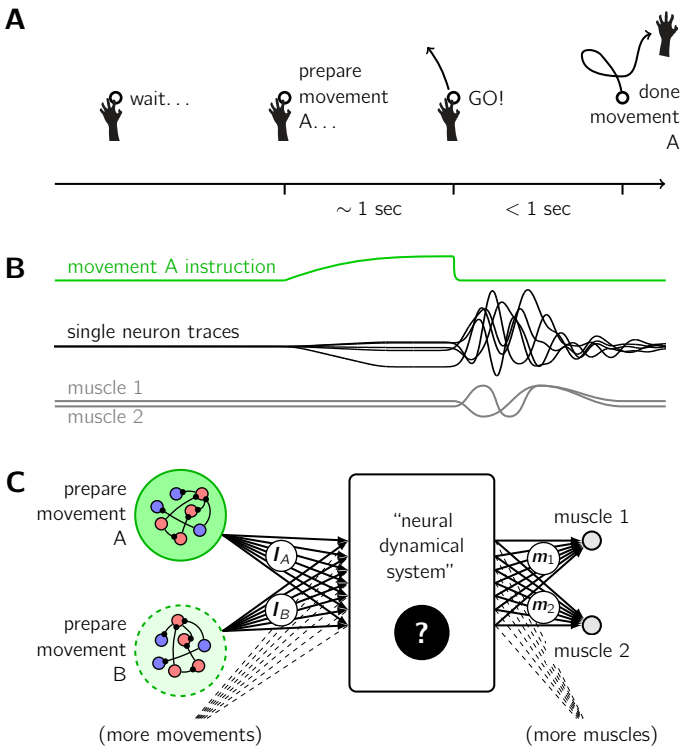


Figure 3.1: **Dynamical systems view of movement planning and execution.** Typical delayed movement generation tasks go as sketched in (**A**). First comes the instruction of what movement must be prepared. The arm must then be held still until the arrival of a go cue, following which the desired arm movement is performed. Muscle activities, which ultimately set the arm in motion, are thought to be read out from a population of motor cortical neurons ("neural dynamical system" in (**C**)). Thus, to generate a certain movement, the dynamical system must produce a specific transient pattern of joint firing activity (see movement-related neuron and muscle activities in (**B**). The goal of the preparatory period is to initialize the neural population in a state that elicits the right population transient. Here, we postulate the existence of movement-specific populations ((**C**), green) that slowly activate following the instruction, and shut off very quickly after the go cue ((**B**), green). These pools feed the motor cortex population through specific sets of weights, such that it is brought to the optimal initial state by the end of the preparatory period ((**B**), black).

3.2 Results

To model the "neural dynamical system" shown in the schematics of Figure 3.1C, we use a conventional network of N interconnected neurons (Dayan and Abbott, 2001; Gerstner and Kistler, 2002b; Miller and Fumarola, 2011), described by a vector $\boldsymbol{x}(t)$ of activation variables which evolve through time according to

$$\tau \frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{W}\, g\left[\boldsymbol{x}(t)\right] + \boldsymbol{\xi}(t) + \sum_{\text{mvt } k} r_k(t)\boldsymbol{I}_k \tag{3.1}$$

Here $g\left[\boldsymbol{x}(t)\right]$ denotes the point-wise application of a saturating nonlinearity, reminiscent of the effective single-neuron f-I curve in a balanced network operating in the asynchronous and irregular firing regime (Methods). Thus, $g\left[x_i(t)\right]$ denotes the momentary firing rate of neuron i, relative to a baseline rate $r_0 = 5\text{Hz}$. In Equation 3.1, τ summarizes the time constants of the single neurons and synaptic dynamics and is set to 200ms to match the dominant timescale in the data of Churchland et al. (2012) (Discussion). $\boldsymbol{\xi}(t)$ is a vector of

N independent, time-varying noisy inputs, and $\boldsymbol{W}$ is the matrix of synaptic connectivity (see below), which plays a dominant role in shaping the dynamics of the network.

In the data of Churchland et al. (2012), firing rates were obtained by averaging spiking activity across many trials for each movement condition. Here, we may think of each "neuron" in our network as a small cluster of neurons that all behave in a similar way, so that $g\,[x_i(t)]$ can readily be interpreted as a trial-averaged firing rate. Consequently, the level of noise (strength of $\boldsymbol{\xi}(\boldsymbol{t})$) is taken to be relatively low, so that firing rate fluctuations during spontaneous activity (before target onset) be as small as they appear in the trial-averaged data (Methods).

To model the preparatory period, we assume that for each instructed movement (movement A, movement B, ...), there is a pool of prefrontal cortical neurons that becomes progressively active during the delay period (Fuster and Alexander, 1971; Amit and Brunel, 1997; Wang, 1999), and feeds the motor network through a fixed set of input weights ($\boldsymbol{I}_A, \boldsymbol{I}_B, \ldots$) (Figure 3.1C). This input is therefore modelled in Equation 3.1 by the last term $\sum_{\text{mvt } k} r_k(t)\boldsymbol{I}_k$, where $r_k(t)$ denotes the temporal activation profile of pool k: either zero if movement k is not to be performed, or the ramp sketched in green in Figure 3.1B if movement k is to be prepared and executed (see also Methods). Thus, during movement preparation, the corresponding "command" pool activates, and brings the system to a steady state from which it is left to evolve freely as the go cue shuts off the command pool. For the preparatory period to initialize the system in state $\boldsymbol{V}_k$, the input weight vector $\boldsymbol{I}_k$ must be set to $\boldsymbol{V}_k - \boldsymbol{W}g\,[\boldsymbol{V}_k]$.

3.2.1 Complex inhibition-stabilized networks (cISNs)

For the network to produce complex patterns of transient firing following the go cue, its connectivity must presumably be equally complex. In situations where the actual network connectivity is not known, random networks are often the default assumption for local microcircuit wiring. Connecting a network randomly does establish complex recurrent pathways, but random networks suffer from the following dilemna: if synaptic efficacies take on weak values, any initial condition $\boldsymbol{V}_k$ will decay roughly exponentially following the go cue, and very little transient amplification is to be expected (Hennequin et al., 2012). If, on the other hand, synaptic connections are strong, such networks exhibit never-ending chaotic activity of large amplitude (Sompolinsky et al., 1988), and the transient aspect of the data is lost (but see Discussion). In the chaotic regime, the mechanism by which the preparatory period could force the system into a desired initial condition is not even clear, and assuming it were possible, the network would not respond reliably to such initial condition in the face of ongoing noise.

We reason that, if there were a way of stabilizing strong random connectivity while preserving

the connection strengths, then reliable transient amplification of well-chosen inputs would become possible. Such networks could then support the generation of fast and complex movements. To test this idea, we engineer connectivity matrices in which the excitatory subnetwork is randomly wired with strong synaptic efficacies, while stability is rescued by an adequate inhibitory feedback loop (Figure 3.2A). We call the result a "complex inhibition-stabilized network", or cISN. Here stability is understood as the local stability of small firing rate fluctuations in the vicinity of the baseline r_0. Therefore, stability is measured by the spectral abscissa $\alpha(\boldsymbol{W})$ of the connectivity matrix. For the network dynamics to be stable around the background state, $\alpha(\boldsymbol{W})$ must be smaller than one.

Technically, cISNs are obtained here from strongly coupled random balanced networks, of which the inhibitory synapses are progressively refined to restore dynamical stability (Figure 3.2). Inhibitory tuning is done following the procedure outlined below and described in more details in the Methods. Briefly, the stabilization procedure implements a gradient descent on the smoothed spectral abscissa (Vanbiervliet et al., 2009), an upper bound on $\alpha(\boldsymbol{W})$ of which the derivatives with respect to the inhibitory synaptic weights can be computed efficiently. The gradient descent operates under three constraints. First, inhibitory weights must remain negative, that is, inhibitory neurons must remain inhibitory. Second, for reasons discussed below, we enforce a global balance of excitation and inhibition by keeping the average inhibitory weight at three times its excitatory counterpart. Finally, to increase the plausibility of the resulting connectivity, the density of inhibitory connections is constrained to a maximum of 40% (Methods). Importantly, although this procedure could perhaps be approximated by a biologically feasible inhibitory plasticity rule (Vogels et al., 2011; Luz and Shamir, 2012), we consider it more conservatively as a principled way of engineering cISNs under constraints, yielding functional circuits that we can further analyze and confront to experimental data. Thus, it is implicitly assumed that the exact procedure by which one (or the brain) arrives at a cISN does not really matter, in other words, that any network that qualifies as a cISN would behave in similar ways as the networks we obtain here (Discussion).

We illustrate the above stabilization procedure on a randomly connected balanced network of size $N = 200$ (100 exc. neurons, 100 inh. neurons) with strong synaptic weights, shown in Figure 3.2C. All connections are initially formed at random with probability $p = 0.1$, non-zero synapses assuming a value of either $+w_0/\sqrt{N}$ or $-3w_0/\sqrt{N}$ depending on the nature of the presynaptic partner. The weight strength w_0 is chosen so as to yield strongly unstable dynamics, with an initial spectral abscissa of 10 (Figure 3.2B). Perhaps surprisingly, stability in such random balanced networks cannot be rescued by merely increasing the overall relative strength of inhibition (Rajan and Abbott, 2006), reflecting the complexity of the recurrent pathways induced by random wiring. Thus, inhibition must be finely tuned in order to stabilize the circuit, which is successfully achieved by our inhibitory tuning procedure

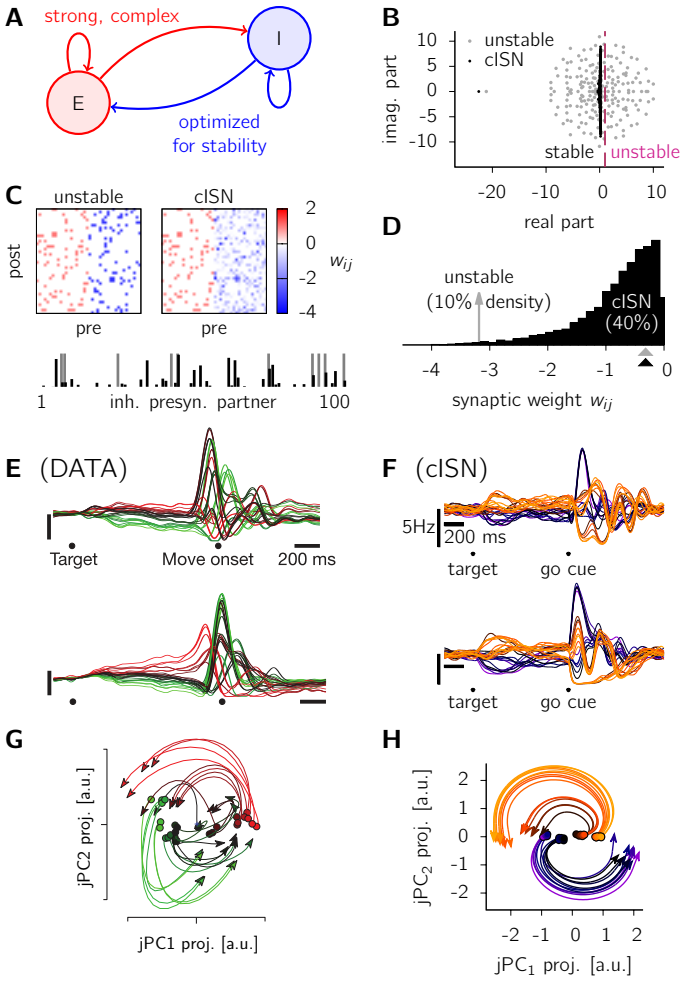


Figure 3.2: **Complex inhibition-stabilized networks (cISNs) match the dynamical behavior of motor populations qualitatively.** (**A**) Schematics of our approach. A population of excitatory neurons is recurrently connected through strong and intricate pathways. This population alone would be dynamically unstable, but is stabilized by fine-tuned inhibitory feedback. (**B**) Eigenvalue spectra of the unstable random balanced network from which the cISN derives (gray), and of the obtained cISN (black). Stability requires all the eigenvalues to lie to the left of the dashed vertical line. Note the large negative real eigenvalue, which corresponds to the spatially uniform activity pattern (the "DC" mode). (**C**) Matrices of synaptic connectivity before (unstable) and after (cISN) stability optimization through inhibitory tuning. Matrices were thinned out to 40×40 (instead of 200×200) for visualization purposes. The bottom row shows the magnitude of all the inhibitory input synapses to a single sample neuron, in the unstable network (gray) and in the corresponding cISN (black). (**D**) Distribution of inhibitory synaptic weights in the unstable network (10% connection density, gray peak at $w_{ij} \simeq -3.18$) and in the stabilized version (40% connection density, black). The mean inhibitory weight is the same before and after optimization ($\simeq -0.318$, gray and black triangle marks). (**E**) Experimental data, adapted from Churchland et al. (2012). Each trace denotes the trial-average firing rate of a single cell (two sample cells are shown here) during a delayed reaching task. Each trace corresponds to one of 27 different reach types (target position / reach curvature). Vertical scale bars denote 20 spikes/sec. The go cue is not explicitly marked here, but occurs about 200ms before movement onset. (**F**) Time-varying firing rates of two neurons in the cISN, for 27 "conditions", each characterized by a different collective steady state of preparatory activity (see text). (**G**) Experimental data adapted from Churchland et al. (2012), showing the first 200ms of movement-related population activity projected onto the top jPC plane. Each trajectory corresponds to one of the 27 conditions mentioned in (E). (**H**) Same analysis as in (G), for the cISN.

(Figure 3.2B). The constrained gradient descent converges, and the spectral abscissa reaches a final value of about $0.18 \ll 1$, indicating that feedback inhibition is properly tuned against the distabilizing effects of the strong excitatory recurrence. The distribution of inhibitory

synaptic strengths in the resulting cISN is wide (Figure 3.2D), and the connectivity looks random in most respects, essentially because the unstable excitatory connectivity is random in the first place (Figure 3.2C). However, shuffling the inhibitory weights destroys stability entirely (not shown).

In the following, we report on this single cISN, but we found it was always possible to stabilize strongly recurrent excitatory networks, provided the feedback loop was given enough "degrees of freedom", i.e. provided there were enough inhibitory synapses to optimize (number of inhibitory neurons, multiplied by the prescribed maximum density of inhibitory connections). Similarly, all cISNs we built operated in a qualitatively similar dynamical regime as the one we are about to report. We also note that the optimization of the connections made by the inhibitory population onto itself is as crucial to stability as the optimization of the connections made onto excitatory cells (something already pointed out by Tsodyks et al. (1997) in their analysis of simpler, 2-dimensional ISNs).

3.2.2 cISNs exhibit complex transient amplification

Now that we know how to build cISNs, we may ask whether they can indeed produce the kind of complex transient behavior that is seen in the data of Churchland and Shenoy (2007) and Churchland et al. (2012) (Figure 3.2E). We find that, provided the preparatory period leaves the system in an appropriate initial condition by the time the go cue arrives, cISNs evoke strong and multiphasic transient firing patterns that match the data qualitatively (Figure 3.2F). In Churchland et al. (2012), the data was obtained by recording the activity of populations of neurons in the motor and premotor cortical areas while a monkey was performing delayed arm movements in a setting similar to the one sketched in Figure 3.1A. For their data shown in Figure 3.2E, there were 27 different reach conditions, defined by various combinations of target position and instructed reach curvature. To implement this task in our model (Figure 3.2F), we assume the presence of a command pool for each of 27 virtual conditions, each pool projecting to the cISN with its own set of input weights. Input weights are chosen so that the preparatory period for each movement initializes the cISN in a state different across conditions but from which large transients are invariably triggered (see below, and Methods).

How must the network be "prepared" for large transients to be elicited as observed in the data? To find the preferred initial conditions of the cISN, we rephrase the problem of collective input tuning as one of energy maximization. We introduce the notion of "evoked energy" $\mathcal{E}(\boldsymbol{a})$, defined as the integrated squared length of the activity vector as the network

evolves freely without input noise from some initial condition $\boldsymbol{x}(t=0) \equiv \boldsymbol{a}$:

$$\mathcal{E}(\boldsymbol{a}) = \frac{2}{\tau \|\boldsymbol{a}\|} \int_0^\infty \|g[\boldsymbol{x}(t)]\|^2 \, dt. \tag{3.2}$$

Here $2/\tau \|\boldsymbol{a}\|$ is a normalizing factor such that $\mathcal{E} = 1$ for an unconnected network ($\boldsymbol{W} = 0$) irrespective of the initial condition $\boldsymbol{a}$. Thus, the energy $\mathcal{E}(\boldsymbol{a})$ measures both the amplitude and duration of the collective transient evoked by initial condition $\boldsymbol{a}$. The problem of finding the initial condition $\boldsymbol{a}$ that evokes maximum energy turns out to have a simple solution in the linear regime, i.e. in situations where the firing rates of the neurons vary in a range over which the slope of their gain function $g[\cdot]$ does not change greatly (Methods). We can even compute a complete basis $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N\}$ of N preferred network states, each of which is successively defined as the initial condition that evokes maximum energy with the constraint that it must be orthogonal to all previous ones. This analysis reveals that, in the linear regime, our cISN transiently amplifies a large set of orthogonal initial conditions (Figure 3.3A). Following initialization in its top preferred initial state $\boldsymbol{a}_1$, the energy it evokes is almost 25 times greater than expected from a mere exponential decay of the initial condition (the default behavior of the individual neurons taken in isolation). The length of the activity vector grows transiently to almost four times its initial length, after which the network returns to a state of baseline firing (Figure 3.3C, top). The return to rest occurs within a time $\simeq 3\tau$, where τ is the intrinsic time constant of a single cell. Note that some cells become transiently more active than baseline, some become less active, and some display multiphasic firing rate patterns. Overall, the population-averaged firing rate remains roughly constant during the transient (red line in Figure 3.3C, middle). A similar behavior, though progressively attenuated, is observed for the top ~ 100 preferred initial states (top ten are shown in Figure 3.3C, upper plot). An equally large number of initial conditions are on the contrary actively suppressed by the recurrent dynamics (Figure 3.3A). For such initializations, the network goes back to baseline firing in a time much shorter than τ, and no transient amplification occurs along the way (Figure 3.3C, bottom). Amplification in the cISN is selective, in the sense that only the first 17 initial states (out of 200 possible orthogonal states) are amplified by a factor greater than $3\mathcal{E}_0 \simeq 11.25$, where $\mathcal{E}_0$ is the energy that a random initial condition is expected to evoke (the average of the energy curve in Figure 3.3A, see the black triangular mark).

The above analysis of the cISN's preferred initial states holds in the linear regime. Clearly, some aspects of the data of Churchland et al. (2012) reveal nonlinear phenomena: for example, many cells see their firing rates quickly decrease down to zero following the go cue, which therefore triggers the lower saturation of their f-I curve (Figure 3.2E). The linear assumption does not hold in the model either, as soon as firing rates during the preparatory period spread over a realistic range. This is precisely because the preferred initial conditions

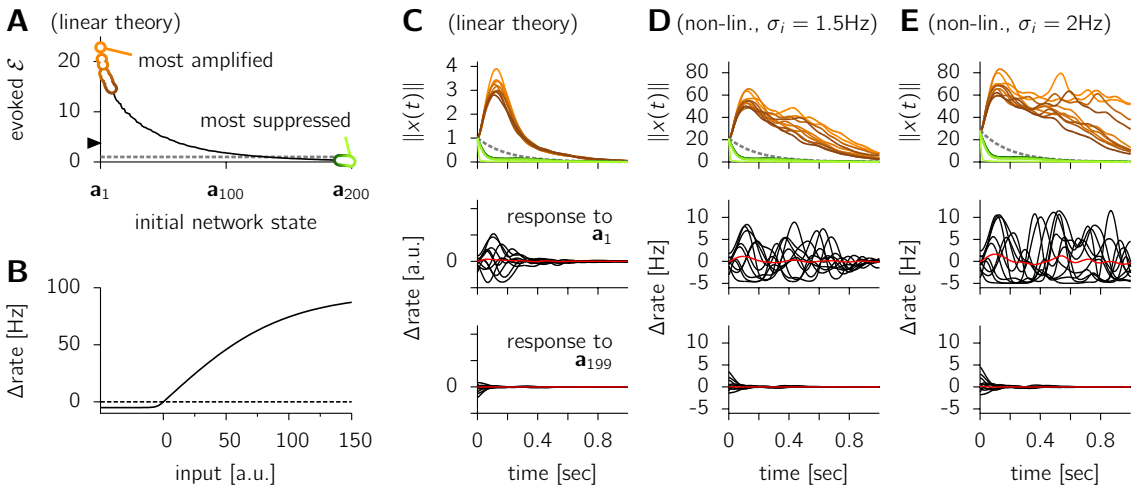


Figure 3.3: **Transient amplification in cISNs** – (**A**) The energy $\mathcal{E}$ evoked by $N = 200$ orthogonal initial conditions $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N\}$ as the network evolves linearly ($g(x) = x$) with no further input ($\boldsymbol{\xi}(t) = 0$) according to Equation 3.1. The energy (Equation 3.2) is normalized such that is equals one for an unconnected network ($\mathbf{W} = 0$) irrespective of the initial condition (dashed horizontal line). Each successive initial condition $\boldsymbol{a}_i$ is defined as the one that evokes maximum energy, within the subspace orthogonal to all previous input patterns $\boldsymbol{a}_{j<i}$ (Methods). The black triangular mark indicates the mean $\mathcal{E}_0$, or the expected evoked energy when the neurons are initialized in random and independent activity states. (**B**) Single-unit input-output nonlinearity ($g[\cdot]$ in Equation 3.1). (**C**) Dynamics of the cISN in the linear regime ($g[x] = x$). Top: time-evolution of the norm $\|x(t)\|$ of the network activity as the dynamics unfold from either of the 10 best or 10 worst initial states (same color code as in panel (A)). The dashed gray line shows $\exp(-t/\tau)$, i.e. the behavior of an unconnected pool of neurons. Bottom: sample firing rate responses of 10 randomly chosen neurons following initialization in state $\boldsymbol{a}_1$ or $\boldsymbol{a}_{199}$. The red line indicates the momentary population-averaged firing rate. (**D,E**) Same as in (C), now with the nonlinear gain function shown in (B). Unlike in the linear case, the dynamics now depend on the spread σ_i of the initial firing rates across the network (1.5Hz in (D), 2Hz in (E)). The larger this spread, the longer the duration of the population transient. When $\sigma_i > 3$, the network initiates self-sustained (chaotic) activity (not shown).

that we search for are indeed amplified, so that some neurons would in principle like to decrease their firing rates by more than the baseline r_0 following the go cue, which is not allowed by the saturating nonlinear gain function (Figure 3.3B). Nevertheless, the onset of amplification is a linear phenomenon, so the above linear analysis of collective tuning provides a very good guess of the network's preferred input patterns in the nonlinear regime (Figure 3.3D and 3.3E).

By design, stability in the cISN is achieved only in the vicinity of the background state, i.e. for relatively small firing rate fluctuations around r_0. When the spread σ_i of the initial firing rates reached by the end of preparatory period is large enough, the network initiates a collective transient similar to what is observed for smaller initial states, but the transient then goes on for much longer durations (Figure 3.3E). In fact, for large initial conditions

($\sigma_i > 3$), the network is apparently able to sustain activity indefinitely, just like an untuned chaotic network would. This dynamical behavior is beyond the scope of this study. Here we calibrate the projection weights of the movement command pools such that, by the end of the preparatory period, the firing rates in the cISN reach a standard deviation of $\sigma_i = 1.5$Hz. In this case, provided the initial state spans the network's top 10 initial conditions or so, complex transient dynamics unfold over only a second or so, which is of the same order as the duration of the movements we consider later.

3.2.3 Rotational collective dynamics in cISNs

Churchland and colleagues reported another important aspect of the transient collective dynamics in motor and premotor cortical areas following the go cue: the complexity of the single-neuron multiphasic responses is in fact hiding orderly rotational dynamics on the population level. That is, they were able to find a plane of projection in which the vector of population firing activity (corresponding to $g[\boldsymbol{x}(t)]$ in our model) would start rotating after the go cue, and rotating consistently in the same direction for all target locations and reach curvatures (Figure 3.2G). This plane was found by applying a dynamical variant of principal component analysis called jPCA to the data (Churchland et al. (2012) – see also Methods and subsection 3.5.3). Surprisingly, such an oscillatory collective behavior is also present in our model, as shown in Figure 3.2H. Just after the go cue, the cISN population activity strongly rotates in the top jPC plane, and rotates consistently in the same direction for all 27 initial conditions previously chosen to mimic the 27 types of arm reaches in Figure 3.2F.

3.2.4 Complex movement generation

The complicated, multiphasic nature of the single-neuron firing rate transients in cISNs suggests the possibility of reading out equally complex patterns of muscle activity. We illustrate this idea in a task where the muscles must produce either of two target movements ("snake" and "butterfly") as depicted in Figure 3.4A. Each movement must be generated from the first 500ms of network dynamics following the go cue. The preparatory input for the "snake" movement is chosen such that, by the arrival of the go cue, the network activity matches the network's preferred initial condition $\boldsymbol{a}_1$. Similarly, planning the "butterfly" movement sets the network in its second preferred initial state $\boldsymbol{a}_2$. Thus, for both movements, the go cue triggers the same kind of transient collective dynamics that we have discussed above (e.g. Figure 3.2F). A single pair of muscle linear readouts is then learned on the basis of 100 noisy trials for both movements (Methods). The two complex trajectories are properly learned (compare the five test trials in Figure 3.4C), although some of the finer details of

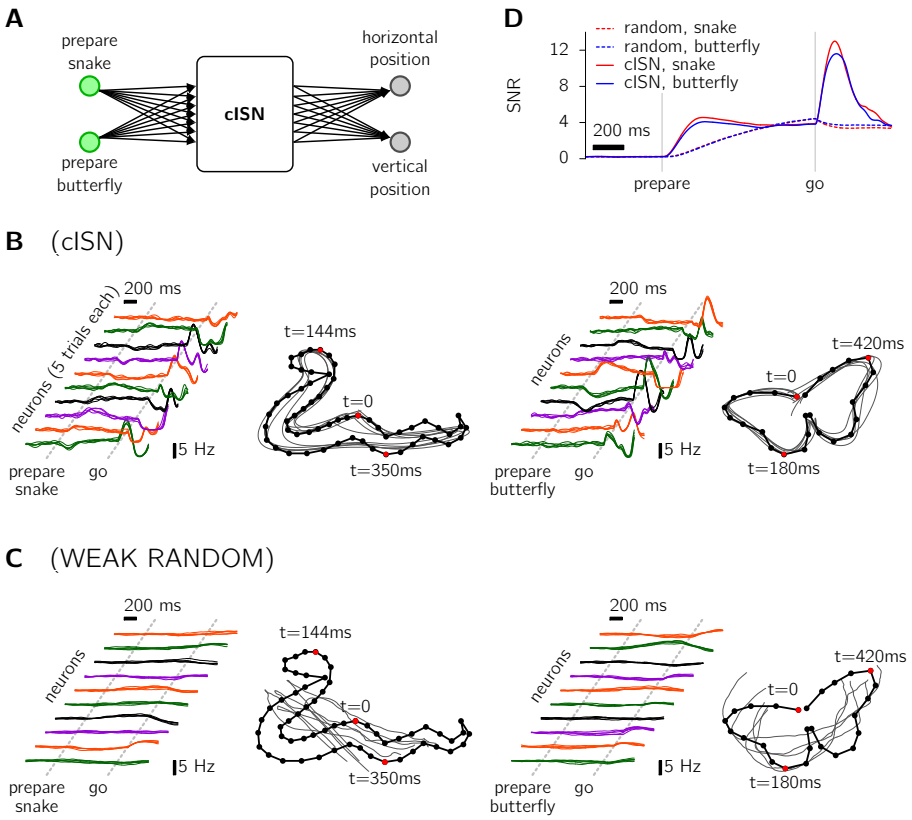


Figure 3.4: **Generation of complex movements through cISN dynamics**. (**A**) Two muscles (gray circles) representing vertical and horizontal positions are activated as linear combinations of the single neuron activities in the cISN during the first 500ms following the go cue. The linear readouts are learned so that the joint activation of the two muscles draws either a snake or a butterfly, depending on which command pool (green circles) activates during the preparatory period. Learning is done on 100 training trials, and we report here on 5 test trials. (**B**) Firing rates versus time for 10 units in the cISN, as the system prepares and executes either of the two target movements. Five trials are shown for each unit. The corresponding muscle trajectories following the go cue are shown for five test trials (gray traces) and compared to the target movement (black). (**C**) Same as in (B), for a weakly connected (untuned) random balanced network. The spectral radius of the corresponding connectivity matrix was set to 0.5. (**D**) Time evolution of the signal-to-noise ratio (SNR) for both movements and both networks (see text).

the target movements (e.g. the fast turns in the snake's tail) cannot be captured owing to the limited number of timescales present in the single-cell firing rate responses.

The quality of the linear readout depends upon two properties. First, it depends on the signal-to-noise ratio (SNR) in the population activity itself. If, on average across trials, the single unit firing rates were to deviate only weakly from baseline during movement-related activity, the magnitude of the population response would be dominated by the recurrent filtering of the noise term $\boldsymbol{\xi}$ in Equation 3.1, which is not movement-specific and varies randomly from

trial to trial. This concern is expressed in our definition of the SNR at time t:

$$\mathrm{SNR}(t) = \sqrt{\frac{\sum_{i=1}^{N} [\mu_i(t) - \mu(t)]^2}{\sum_{i=1}^{N} \langle \varepsilon_i(t)^2 \rangle}} \tag{3.3}$$

where $\langle \cdot \rangle$ denotes trial averaging, $\mu_i(t) = \langle g[x_i(t)] \rangle$ is the trial-averaged firing rate of unit i at time t (the "signal"), $\mu(t) = \sum_i \mu_i(t)/N$, and $\varepsilon_i(t) = g[x_i(t)] - \mu_i(t)$ is the noise present in unit i. Thus, the signal is measured by the spread of the trial-averaged momentary firing rates across the population, while the noise is given by the average variance of the trial-to-trial rate fluctuations in single units. In our model, as long as the dynamics remain approximately linear, the noise term (denominator in Equation 3.3) is essentially set by the amplitude of the noisy inputs $\boldsymbol{\xi}(t)$ in Equation 3.1. The signal power (numerator), on the other hand, critically depends on the selective amplification of the preparatory states by the recurrent dynamics. Here the preparatory states for the "snake" and "butterfly" movements are purposely chosen as the top two preferred initial conditions ($\boldsymbol{a}_1$ and $\boldsymbol{a}_2$) of the cISN. As a result, the trial-averaged firing rates in the network transiently expand around the baseline rate following the go cue (Figure 3.4B). Accordingly, the SNR transiently increases by a factor of 3 or more (Figure 3.4D). The quality of the readout is also influenced by the magnitude of the readout weights found by the linear regression (relative to the amplitude of the movement itself). The noisy trial-to-trial fluctuations of the network activity introduce random errors in the muscle trajectories, with a variance proportional to the squared norm of the optimal readout weights. Large weights imply that muscle activities are obtained from cancellations between large positive activities in some neurons and large negative activities in some other neurons (this situation is more commonly referred to as overfitting). Such solutions are likely to arise when the network dynamics are not rich enough in comparison with the complexity of the desired movements. Here, the cISN generates complex patterns of firing: neurons display asynchronous, multiphasic firing patterns that form a rich set of basis functions (Figure 3.4B) from which complex movements may be decoded.

Weakly connected random balanced networks are one example of networks that have none of the above two features: they do not act as strong selective amplifiers, and the activity transients they elicit are close to simple exponential decays (Figure 3.4C). We test such a network (randomly connected balanced network with a spectral abscissa of $R = 0.5$) on the same movement generation task. The quality of the decoded movements is much poorer, and more variable from trial to trial (compare the five test trials in Figure 3.4C).

3.2.5 Balanced amplification

We have seen that the cISN strongly but transiently amplifies certain "preferred" network states. The network activity after 200ms of recurrent processing has much larger amplitude than, but bears little spatial resemblance to, the initial condition. This is reflected in the fast decay of the correlation coefficient between the network activity and the initial state (Figure 3.5A, black). Interestingly, the dynamics of this correlation are not the same, however, when the excitatory and inhibitory sub-populations are considered separately. The inhibitory activity becomes very quickly negatively correlated with its initial state, while the excitatory activity remains positively correlated for the entire duration of the transient (compare the red and blue curves in Figure 3.5A). This indicates that, during the course of amplification, the spatial pattern of excitatory activity is amplified but does not change much, while that of inhibitory activity quickly reverses (i.e. changes sign) in order to quench the excitatory transient and pull the system back to rest. Such dynamics are reminiscent of "balanced amplification", a mechanism previously described by Murphy and Miller (2009) to account for the spontaneous emergence of structured activity patterns in a model of primary visual cortex. Balanced amplification refers to the transient amplification of "difference modes" in which the excitatory and inhibitory sub-populations fire in spatially equal but opposite ways, into "sum modes" in which the activity of both populations equalize. Here, our cISN is not connected following any clear topology. Thus, it is difficult to define sum/difference modes as in Murphy and Miller (2009), that is, as spatial patterns of balance/imbalance in firing activity. However, it is possible to define sum/difference modes as patterns of balance/imbalance in the excitatory and inhibitory *inputs* across the population. Figure 3.5C shows the dynamics of the Pearson correlation coefficient r_{EI} between the vector of momentary excitatory inputs in the network (the product of the $N \times N/2$ excitatory sub-matrix of $\boldsymbol{W}$ with the vector of excitatory firing rates at time t), and its inhibitory counterpart. The preferred initial states tend to break the balance of excitatory and inhibitory inputs, yielding significantly negative correlations between these two input vectors ("difference" modes). The input balance is quickly restored by the recurrent dynamics ("sum" modes), with a correlation topping ~ 0.8 after 160ms following the first preferred initial condition. Sample pairs of excitatory and inhibitory input currents are shown in Figure 3.5B for two units to illustrate the effect.

As in Murphy and Miller (2009), amplification in the cISN relies on the connectivity matrix $\boldsymbol{W}$ being mathematically "nonnormal" (the eigenvectors of $\boldsymbol{W}$ do not form an orthogonal basis). The above analysis in terms of sum/difference modes suggests that the connectivity in the cISN may be functionally equivalent to a set of feedforward links from difference modes to sum modes. Murphy and Miller (2009) used a Schur decomposition of $\boldsymbol{W}$ to reveal these feedforward connections and to draw a simple picture of the dynamics. This was made

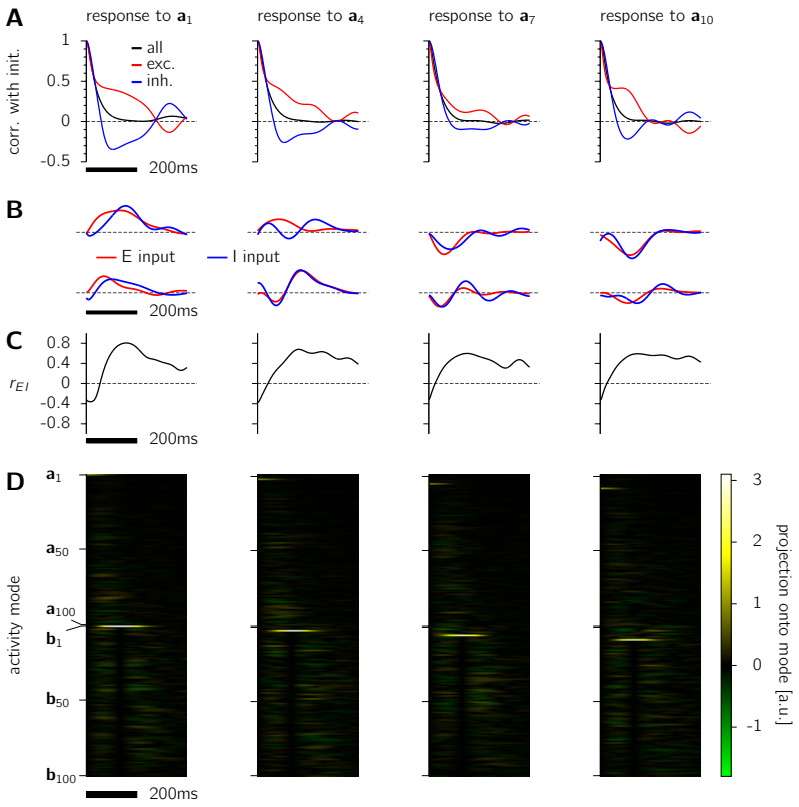


Figure 3.5: **Balanced amplification in cISNs**. The network is initialized in either of its first, fourth, seventh or tenth preferred initial states, from which the recurrent dynamics run freely with no further inputs. The amplitude of the initial condition is chosen weak enough for the dynamics of amplification to remain linear (c.f. Figure 3.3). In all panels here, the first 400ms of dynamics are shown. (**A**) Time course of the Pearson correlation coefficient between the network activity (black) and the initial state. The red (resp. blue) line denotes the same analysis, restricted to the activity of the excitatory (resp. inhibitory) neurons. (**B**) Time course of the excitatory (red) and inhibitory (blue) inputs for two sample units. (**C**) Time course of the Pearson correlation coefficient r_{EI} between the vectors of momentary excitatory and inhibitory inputs to the N single units. The preferred initial states tend to break the input balance. The recurrent dynamics then restore the excitation-inhibition balance in less than 200ms. (**D**) Network activity expressed in a basis of orthogonal activity modes $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{100}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_{100}\}$ that yields a parsimonious representation (see text). Colors represent the projection of the population activity onto the corresponding mode.

possible by the topological regularities of their V1 model connectivity. Here we take an alternative approach to finding a basis of orthogonal vectors in which the network activity takes a simple form. We know that the strongest feedforward links should originate mostly from the top preferred initial states of the network, already found above (c.f. Figure 3.3). We thus take the first $N/2 = 100$ top preferred initial states to form the first half $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{100}\}$ of the orthogonal basis. To complete the basis, we take the network's response to each of these initial conditions at time $t = 2\tau/3 \simeq 133$ms (roughly the time at which the network response has largest amplitude), resulting in $\{\boldsymbol{b}'_1, \boldsymbol{b}'_2 \ldots, \boldsymbol{b}'_{100}\}$. We then orthonormalize the $\boldsymbol{b}'_i$s against the $\boldsymbol{a}_i$s, and obtain a complete orthogonal basis $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{100}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_{100}\}$ in which we hope the network dynamics will have a simple form. The projection of the network activity onto those basis vectors is depicted in Figure 3.5D, following initialization in each of the first, fourth, seventh and tenth preferred network states. It becomes apparent that

the network's response to $\boldsymbol{a}_1$ is dominated by mode $\boldsymbol{b}_1$, and similarly for the subsequent pairs $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ for which significant energy is evoked[1]. Note that since all the $\boldsymbol{b}_i$s are orthogonal to one another, the various responses to the top preferred initial states are therefore highly distinguishable. This explains why the same network could elicit two very different muscle trajectories provided the network was initialized in different preferred states for each movement (Figure 3.4B).

Were the cISN connectivity to embed only pairwise feedforward links of the form $\boldsymbol{a}_i \rightarrow \boldsymbol{b}_i$ between activity modes, single-cell responses would not look multiphasic as in Figure 3.3C. Indeed, $\boldsymbol{a}_i$ would only decay, sourcing $\boldsymbol{b}_i$ which would thus rise and decay (Murphy and Miller, 2009), and such a monophasic transient would show up in the single-neuron responses too. Looking more closely at the details of Figure 3.5D, one may see that following initial state $\boldsymbol{a}_i$, a significant amount of energy is also developed and distributed along the other $\boldsymbol{b}_j$s $(j \neq i)$. Those responses are mostly biphasic (see the fine details of the heat maps in Figure 3.5D), and interfere very little with the response in $\boldsymbol{b}_i$ at time $2\tau/3$. Thus, although the network response at the peak of amplification is very well described in terms of pairwise balanced amplification links, what happens before and after is more complicated and reflects the complex nature of the excitatory connectivity that cISNs are obtained from.

3.2.6 Structure of spontaneous activity in cISNs

We now look at the structure of spontaneous activity in cISNs. Here we define "spontaneous" activity as the network activity in the absence of a specific stimulus, i.e. when the inputs to each neuron are restricted to i) a private source of noise ($\boldsymbol{\xi}(t)$ in Equation 3.1) and ii) the recurrent synaptic input. The external noise being independent across neurons, it randomly stimulates each of the orthogonal activity modes $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N\}$ with equal intensity, and so long as the dynamics remain reasonably linear we expect the network's responses to those input fluctuations to superimpose. Since, on average, the $\boldsymbol{a}_i$ input patterns are amplified by the recurrent circuitry (Figure 3.3A), single-unit spontaneous activity fluctuations are larger in the cISN than they would be in an unconnected network with identical input statistics (Figure 3.6A). Moreover, because amplification is selective, spontaneous activity fluctuations in the cISN are expected to be comparatively larger along those modes $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots$ that are best amplified (c.f. Figure 3.5D). This is likely to shape the structure of the spontaneous pairwise correlations in the cISN. For example, neurons that are jointly and strongly active in the most amplified activity pattern $\boldsymbol{b}_1$ are likely to have positively correlated spontaneous fluctuations.

[1] In fact, the basis we have just derived is an approximate Schur basis for the matrix $\exp[2(\boldsymbol{W} - \mathbb{1})/3]$, which maps the initial condition onto the network response after $2\tau/3$ seconds of recurrent dynamics. That matrix becomes lower-triangular in the new basis, with all significant entries lying along the secondary diagonal (i.e. $t_{ij} \neq 0$ for $i = j + N/2$).

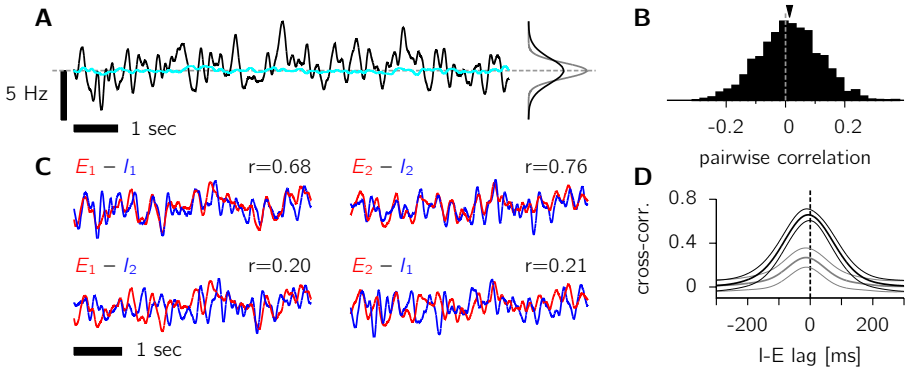


Figure 3.6: **In cISNs, excitatory and inhibitory inputs to single cells are precisely matched during spontaneous activity.** (**A**) Black: spontaneous fluctuations around baseline rate ($r_0 = 5Hz$, dashed horizontal line) of a sample cell in the network. The corresponding rate distribution is shown on the right (black), and compared to the distribution obtained if the cell were not connected to the rest of the network (gray). The cyan line denotes the momentary population average rate, which fluctuates much less. The strength of the input noise is chosen such that firing rates remain in a range where the gain function is approximately linear. (**B**) Histogram of pairwise correlations measured on 100 seconds of spontaneous activity. The black triangular mark indicates the mean (~ 0.014). (**C**) Excitatory (red) and inhibitory (blue) inputs to two sample cells, normalized to z-score. These are compared in four different ways, and the corresponding Pearson correlation coefficient (over 100 seconds of activity) is indicated above each combination. (**D**) Black: cross-correlogram of excitatory and inhibitory inputs to single cells, each normalized to z-score (cf. (C), top row). The solid line is an average across all neurons; flanking lines denote $\pm$ 1 standard deviation. Inhibition lags behind excitation by a few milliseconds. Cross-correlating the E input into one cell with the I input into another cell (cf. (C), bottom row) yields the gray curve, which is an average over 1,000 randomly chosen such pairs in the cISN.

Many pairs of neurons in the cISN are noticeably correlated (positively as well as negatively), as seen from the wide distribution of spontaneous pairwise correlations in Figure 3.6B. This distribution has a small but significant positive mean ($\sim$ 0.014) which reflects the small fluctuations of the population-averaged momentary firing rate (cyan line in Figure 3.6A), to which – by definition – all cells contribute.

The mechanism of balanced amplification described above (3.2.5) has an interesting consequence on the spontaneous activity: excitatory and inhibitory inputs to single cells are highly temporally correlated (Figure 3.6C). This is because the network quickly amplifies small patterns of imbalance in those inputs into large activity transients that re-establish the balance (Figure 3.5B and 3.5C). At any time thus, the vector of excitatory inputs across the network is to be highly correlated with its inhibitory counterpart. Since spontaneous activity here is stationary, the average momentary balance of E and I inputs across the network translates into an average temporal balance of those inputs in the single units (Figure 3.6C, top row). The Pearson correlation between excitatory and inhibitory synaptic input streams

averages to ~ 0.66 across cells. Furthermore, we have seen that it is mostly the spatial pattern of inhibitory activity that reverses during the course of amplification to restore the balance, while the excitatory activity is much less affected (Figure 3.5A). Thus, inhibitory inputs are expected to lag behind excitatory inputs by a few milliseconds during spontaneous activity, which is indeed the case as shown by the shift in their average cross-correlogram (Figure 3.6D).

That excitatory and inhibitory inputs should be somewhat correlated during spontaneous activity is a known fact in the theory of balanced cortical circuits (Renart et al., 2010), and the non-zero fluctuations of the population-averaged firing rate are an obvious source of such input correlations (see 3.5.2 of the Supplemental Data). Here, interestingly, excitatory and inhibitory inputs are correlated more strongly than expected from the magnitude of the shared population fluctuations. This can be seen by correlating the excitatory input stream taken in one cell and the inhibitory input stream taken in another cell (Figure 3.6C, bottom row). Such correlations average to only ~ 0.26 (to be compared with 0.66 above – see Figure 3.6D). We return to this in the Discussion.

3.3 Discussion

"In what regime does the cortical circuit operate"? Ozeki et al. (2009) recently raised this question, and partially answered it for the cat visual cortex (V1). They argued that V1 operates as a network that i) would be dynamically unstable in the absence of inhibitory feedback, and ii) is successfully stabilized by inhibition. They called a network so defined an "ISN", for "Inhibition-Stabilized Network" (see also Tsodyks et al. (1997)), and found that ISNs explain most known aspects of the suppression of visual cortical responses by appropriate stimulation of the receptive field surround. They were also able to verify a few stringent experimental predictions made by ISN dynamics. Here we have pursued the idea that local microcircuits, as opposed to larger cortical areas such as V1, may also operate as ISNs, despite seemingly unstructured motifs of synaptic wiring. We have introduced the concept of "complex ISNs" to broadly define balanced networks in which the recurrent excitatory connectivity is intricate, strongly unstable on its own, but stabilized by an appropriate inhibitory feedback loop. To study their dynamics, we have provided a principled way of instantiating networks of this new class, through progressive and optimal refinement of the inhibitory synaptic connectivity.

We have found cISNs capable of selective transient amplification of specific input patterns. The single-neuron as well as the collective dynamics in cISNs are in good qualitative agreement with the response properties of motor and premotor cortical neurons during arm reaching (Churchland et al., 2012). Elaborating on the putative functional role of such complex

transients, we have found that cISNs can be used as motor "engines" to generate complicated and reliable movements. Reading out muscle trajectories from the cISN population activity does not require elaborate decoding schemes: simple linear readouts are good enough. Interestingly, the generation of noise-robust, complex movements requires forcing the cISN into either of a few specific preparatory states through the delivery of appropriate inputs, which are then withdrawn to release the network into free movement-generating dynamics. A qualitatively similar phenomenon seems to take place in the data too: (pre)motor cortical circuits first engage into slow preparatory dynamics, settling into some steady-state activity prior to eliciting the movement-related amplifying transients (Churchland et al., 2010a, 2012).

Although the dynamics of cISNs are strikingly similar to those of motor and premotor cortical populations during reaching, and although complex movements may indeed be read out from cISN population transients, cISNs by themselves do not constitute a complete model of movement generation. Several key aspects have clearly been ignored here, the first of which is the control of movement speed and duration that are mostly set here by the intrinsic time constant of the single neurons in the cISN. Likewise, we have completely left aside the problem of active movement *control*, by restricting our study to an open-loop system: the cISN does what it does following the initial condition, the movement is then merely read out from the instantaneous population activity, and there is no mechanism to correct for large external disturbances that may come in after the go cue. Proper movement control undoubtedly requires output signals – such as a visual appreciation of the actual arm position – to be fed back into the system. Nevertheless, the control-theoretic method we have used here to stabilize the network (minimization of the smoothed spectral abscissa) is likely an interesting tool to be used in future studies of feedback-mediated control in neuronal networks (see also Sussillo and Abbott (2009)).

The generation of large transients in cISNs relies primarily on the mechanism of "balanced amplification", first described by Murphy and Miller (2009) in a model of V1 synaptic organization. The authors argued that, in networks with strong excitation balanced by equally strong (or stronger) inhibition, small patterns of spatial imbalance (difference modes) should drive large activity transients in which neighboring excitatory and inhibitory neurons fire hand in hand (sum modes). The absence of a topology in cISNs makes it impossible to tell which neuron is a neighbor to which, thus sum and difference modes are difficult to define. Nevertheless, we have shown that sum/difference modes may alternatively be defined as balance/imbalance in the patterns of excitatory and inhibitory synaptic *inputs* across the network. With such definitions, we have shown that balanced amplification – which we continue to define as the consequence of strong, pairwise feedforward links from difference to sum modes hidden in the connectivity – largely contributes to the dynamics of cISNs.

cISNs capture a key experimental observation regarding how excitation and inhibition interact: during spontaneous activity, balanced amplification of noisy external inputs establishes an exquisite temporal balance of excitatory (E) and inhibitory (I) inputs to single cells. This phenomenon has been observed in several brain areas, and on levels as different as the trial-averaged E and I synaptic input conductances in response to sensory stimuli (Wehr and Zador, 2003; Mariño et al., 2005; Froemke et al., 2007; Dorrn et al., 2010), single-trial synaptic responses in which the trial-average has been removed ("residuals", Cafaro and Rieke (2010)), and spontaneous activity (Okun and Lampl, 2008; Cafaro and Rieke, 2010). In cISNs, the detailed balance emerges from the simultaneous inhibitory anihilation of many distabilizing excitatory pathways initially present in the circuit, which goes beyond simply stabilizing the overall population activity. Here, a word of caution must be given: how should we interpret the fine temporal balance of excitatory and inhibitory inputs reported in experiments? The most obvious source of E/I input correlations are the joint fluctuations of the entire local pool of neurons, E and I included. Indeed, consider a single cell that receives hundreds of E and I synaptic inputs from the local network. Assuming that the E presynaptic partners have only weakly correlated firing rate fluctuations, the compound E conductance that this cell receives is effectively a measure of the average excitatory activity $\mu_E(t)$. Similarly, the compound I conductance is a measure of the average inhibitory activity $\mu_I(t)$. Now, it turns out the strongest balanced amplification link is made by the spatially uniform difference mode $(1, \ldots, 1, -1, \ldots, -1)$ onto the spatially uniform sum mode $(1, 1, \ldots, 1)$. This is true in any balanced network made of separate populations of E and I neurons, irrespective of the details of the connectivity (Murphy and Miller, 2009; Hennequin et al., 2012; Kriener et al., 2008). Thus, unless the sum mode also exerts a strong negative feedback onto itself (see below), the sum mode is driven into large fluctuations during spontaneous activity, causing co-variations in $\mu_E(t)$ and $\mu_I(t)$, and thus co-variations in the E and I input conductances (see also 3.5.2 in the Supplemental Data). Here though, we have found that E/I input correlations in the cISN are 2.5 times greater than expected from the above argument (Figure 3.6D). How can this be? The answer lies in the nature of the pairwise balanced amplification links discussed above: cISNs strongly amplify patterns of imbalance in the E and I inputs across the network, into large patterns of balanced inputs. Importantly, the corresponding activity patterns are centered, i.e. do not interfere with the DC mode. Thus, E/I input correlations may not only emerge from large DC fluctuations, but also from the spontaneous amplification of a collection of centered modes that induce balanced E and I inputs. In which proportions do each of the above two sources of correlations contribute to the balance of input conductances in the cortex? The earliest evidence for the input E/I co-tuning during spontaneous activity actually came from paired recordings in *different*, neighbouring cells (Okun and Lampl, 2008). Thus, by experimental design, only E/I correlations originating from large DC fluctuations was being searched for and could be found. We do not know how greater the measured

E/I input correlation coefficient would have been, had the E/I conductance been recorded simultaneously in single cells. Conversely, a recent study in the mouse retina was able to measure E and I input conductances (near)-simultaneously in single cells, but did not perform cross-measurements in pairs of different cells (Cafaro and Rieke, 2010). Here again, we have no way of estimating the relative contributions of both alternatives. Our results suggest that, in order to be more conclusive, later experiments should ideally attempt to measure the correlation coefficient of E and I inputs both to single cells and to pairs of different cells in the local microcircuit. Should the latter be greater than the former, cISNs dynamics would be a plausible candidate circuit structure, and balanced amplification a plausible mechanism.

Here, we have used an optimal inhibitory stabilization algorithm to instantiate networks from the (potentially large) class of cISNs. Although we have been careful to constrain the optimization procedure to yield plausible network connectivities, our update rule for inhibitory synapses is not (and was not primarily meant to be) a plausible synaptic plasticity mechanism. Indeed, the prescribed synaptic modifications are not readily expressed as functions of pre- and post-synaptic activities. Thus, it would be interesting to relate cISNs (and optimal network stabilization) to recent models of inhibitory synaptic plasticity (Vogels et al., 2011; Luz and Shamir, 2012; Kullmann et al., 2012). In Vogels et al. (2011) (see the last figure of their paper), memory patterns were stored as strengthened connections among pools of excitatory neurons. These connections were strong enough to destabilize the background state, where all network neurons would usually fire at low rate. A simple inhibitory plasticity rule was then shown to restore stability, despite the strengthened excitatory connections being left untouched (as in our study). The memory patterns could then be transiently recalled by breaking the balance of excitation and inhibition in (some part of) the corresponding cell assembly. The network would go back to its background state of low activity as soon as the stimulation was switched off. Such transient amplification behavior shows all the defining characteristics of balanced amplification in (c)ISNs, which suggests interesting ties may exist between their local inhibitory plasticity rule and our optimal stabilization procedure.

3.4 Methods

3.4.1 Network setup and dynamics

The network dynamics are given by Equation 3.1, which we integrate using a standard fourth-order Runge-Kutta method. The noise term $\boldsymbol{\xi}(t)$ is modelled as a collection of N independent Ornstein-Uhlenbeck processes with time constant $\tau_\xi = 50\text{ms}$ (N is the network size). The variance of those processes is set to $\sigma_0^2(\tau + \tau_\xi)/\tau_\xi$ such that, in the limit of very

weak synaptic connectivity, the firing rate of each cell in the network would fluctuate with a standard deviation $\sigma_0 = 0.2$Hz. Following Rajan et al. (2010), we choose the gain function as

$$g(x) = \begin{cases} r_0 \tanh\left(\dfrac{x}{r_0}\right) & \text{if} \quad x < 0 \\ (r_{max} - r_0)\tanh\left(\dfrac{x}{r_{max} - r_0}\right) & \text{if} \quad x \geq 0 \end{cases} \tag{3.4}$$

with $r_0 = 5$Hz and $r_{max} = 100$Hz, so that firing rates effectively vary between 0 and 100Hz, with a 5Hz baseline.

During the preparatory period, the "command" pool corresponding to the desired movement slowly activates with the following continuous temporal profile: slow exponential rise with time constant 400ms from target onset to go cue, then fast exponential decay with time constant 2ms from the go cue on. The overall scaling of the projection weights $\boldsymbol{I}_k$ onto the motor circuit is chosen such that, in the limit of very long preparatory periods, preparatory activity in the network reaches a steady state of standard deviation $\sigma_i = 1.5$Hz.

In Figure 3.2E, 27 conditions are associated with 27 command pools, each projecting onto the motor circuit (cISN) via different weight vectors $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{27}$. We choose the $\boldsymbol{I}_k$ so that the steady state preparatory activity $\boldsymbol{V}_k$ in the motor circuit lies within the subspace spanned by the top two preferred initial conditions of the cISN, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ (see below, 3.4.3). More precisely, projection vectors are chosen as $\boldsymbol{I}_k = \boldsymbol{V}_k - \boldsymbol{W}g[\boldsymbol{V}_k]$ with $\boldsymbol{V}_k = \sum_{\ell=1,2} s_{k\ell} z_{k\ell} \boldsymbol{a}_\ell$ where $s_{k\ell}$ is a random sign, and $z_{k\ell}$ is drawn uniformly between 0.5 and 1.

3.4.2 Connectivity matrices

Random connectivity matrices of size $N = 2M$ (M positive (excitatory) columns and M negative (inhibitory) columns) are generated as in Hennequin et al. (2012), with connectivity density $p = 0.1$. Non-zero weights are set to $\pm w_0/\sqrt{N}$, with $w_0 = R/\sqrt{p(1-p)}$ and a sign that depends on the nature of the presynaptic neuron (E or I). Here R is the desired spectral abscissa (before stability optimization). We then enforce a global balance in favor of inhibition, by writing the connectivity matrix $\boldsymbol{W}$ block-wise as

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}^{E\to E} & \boldsymbol{W}^{I\to E} \\ \boldsymbol{W}^{E\to I} & \boldsymbol{W}^{I\to I} \end{pmatrix}, \tag{3.5}$$

and multiplicatively rescaling both inhibitory blocks (separately) to achieve

$$\overline{\boldsymbol{W}}^{I\to E} = -\gamma \overline{\boldsymbol{W}}^{E\to E} \qquad \text{and} \qquad \overline{\boldsymbol{W}}^{I\to I} = -\gamma \overline{\boldsymbol{W}}^{E\to I} \tag{3.6}$$

where $\overline{\boldsymbol{W}}^{X\to Y}$ denotes the average over all matrix elements and $\gamma = 3$ sets the overall strength of inhibition relative to excitation. For large networks, a value of γ greater than one

ensures that the overall population firing rate remains almost constant in time (Hennequin et al., 2012).

3.4.3 Preferred initial states

The analysis of collective input tuning is done in the linear regime, i.e. assuming that firing rates do not deviate too much from their baseline r_0 so that $g(x) \simeq x$. Let $\boldsymbol{W}$ be a stable connectivity matrix. Imagine initializing the network in a state $\boldsymbol{x}(t=0) = \boldsymbol{a}$ of unit norm and letting the noiseless dynamics run freely according to Equation 3.1 (this corresponds to running Equation 3.1 with $\boldsymbol{\xi}(t) = 0$ and $\boldsymbol{I}(t) = \boldsymbol{a}\,\delta(t)$). We define the evoked "energy" as

$$\mathcal{E}(\boldsymbol{a}) = \frac{2}{\tau}\int_0^\infty \|\boldsymbol{x}(t)\|^2\, dt. \tag{3.7}$$

Here $2/\tau$ is a normalizing factor such that $\mathcal{E} = 1$ for an unconnected network ($\boldsymbol{W} = 0$), irrespective of the unit-norm initial condition $\boldsymbol{a}$ (Equation 3.1 would then give $\|x(t)\|^2 = \exp(-2t/\tau)$). Note that for a stable network, $\mathcal{E}$ is finite, in the sense that any initial condition is bound to decay exponentially, asymptotically in time. We then define the "best" input direction as the initial condition $\boldsymbol{a}_1$ that maximizes $\mathcal{E}(\boldsymbol{a})$. Equation 3.7 can be rewritten as

$$\mathcal{E}(\boldsymbol{a}) = \boldsymbol{a}^T \left[2\int_0^\infty e^{t(\boldsymbol{W}-\mathbb{1})^T}\, e^{t(\boldsymbol{W}-\mathbb{1})}\, dt\right] \boldsymbol{a} \overset{\text{def}}{=} \boldsymbol{a}^T\, \boldsymbol{Q}\, \boldsymbol{a} \tag{3.8}$$

where $(\cdot)^T$ denotes the matrix transpose. The last equality defines $\boldsymbol{Q}$ as the matrix integral inside square brackets. $\boldsymbol{Q}$ is a symmetric, positive-definite matrix, and its principal eigenvector is precisely the initial condition $\boldsymbol{a}_1$ that maximizes the evoked energy, which is then given by the corresponding principal eigenvalue of $\boldsymbol{Q}$. In fact, the full eigenbasis of $\boldsymbol{Q}$, ranked in decreasing order of the associated eigenvalues, defines a collection $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N)$ of N orthogonal input states that each maximize the evoked energy within the subspace orthogonal to all previous best input directions. Again, the eigenvalues are the corresponding evoked energies. We use this energy formalism again below to explain the optimal stabilization algorithm. Note that matrix $\boldsymbol{Q}$ is the solution to the Lyapunov equation

$$(\boldsymbol{W}-\mathbb{1})^T\, \boldsymbol{Q} + \boldsymbol{Q}\,(\boldsymbol{W}-\mathbb{1}) = -2\cdot\mathbb{1} \tag{3.9}$$

which is easily solved numerically (e.g. using the Matlab function `lyap`, see also Bartels and Stewart (1972)). There is therefore no need to actually compute and summate the matrix exponentials that show up in the definition of $\boldsymbol{Q}$ (Equation 3.8).

3.4.4 Optimal inhibitory stabilization

Optimizing a recurrent system for linear stability is a standard problem in control theory. The main difficulty lies in the nature of the spectral abscissa $\alpha(\boldsymbol{W})$ (the most natural objective function to consider for linear stability), which is typically a non-smooth function of matrix $\boldsymbol{W}$, precluding the use of gradient-based optimization methods. Progress on this issue has been made recently through the introduction of the "smoothed spectral abscissa", a relaxation of $\alpha(\boldsymbol{W})$ that is smooth with respect to small variations in $\boldsymbol{W}$ (Vanbiervliet et al., 2009). To introduce this stability measure, let us first recall the expression for the energy evoked in a stable linear network as it evolves freely according to Equation 3.1 from some initial condition $\|\boldsymbol{a}\| = 1$:

$$\mathcal{E}(\boldsymbol{W}, \boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{Q}(1) \boldsymbol{a} \tag{3.10}$$

where $\boldsymbol{Q}(1)$ is defined more generally as

$$\boldsymbol{Q}(s) = 2 \int_0^\infty e^{t(\boldsymbol{W}-s\mathbb{1})^T} \, e^{t(\boldsymbol{W}-s\mathbb{1})} \, dt \tag{3.11}$$

For the network to be stable, the energy must remain finite for any initial condition $\boldsymbol{a}$. Note that $\boldsymbol{Q}(1)$ is a symmetric positive definite matrix, whose leading eigenvalue corresponds to the maximum energy that a unit-norm initial condition can evoke. Thus, $\mathcal{E}(\boldsymbol{W}, \boldsymbol{a})$ is upper-bounded by the largest eigenvalue of $\boldsymbol{Q}(1)$, and since all the eigenvalues are positive, it is also upper-bounded by their sum, i.e. by the trace of $\boldsymbol{Q}(1)$. Thus, if $\mathrm{tr}\,[\boldsymbol{Q}(1)] < \epsilon^{-1}$ for some given $\epsilon > 0$, then the energy evoked by any $\boldsymbol{a}$ is less than ϵ^{-1}, so the network dynamics of Equation 3.1 are guaranteed to be stable. In a network that is not (yet) linearly stable, one can ask: how far to the left must the system be "shifted", $\boldsymbol{W} \mapsto \boldsymbol{W} - s\mathbb{1}$, for $\mathrm{tr}\,[\boldsymbol{Q}(s)]$ to become smaller than ϵ^{-1}? The ϵ-smoothed spectral abscissa is the answer to this question (see supplementary Figure 3.8). Mathematically, $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ is the unique root of $s \mapsto \mathrm{tr}\,[\boldsymbol{Q}(s)] - \epsilon^{-1}$, which is a monotonically decreasing function of s. If the shift s is smaller than the spectral abscissa $\alpha(\boldsymbol{W})$, some of the eigenvalues of $\boldsymbol{W} - s\mathbb{1}$ will have positive real parts, causing $\mathrm{tr}\,[\boldsymbol{Q}(s)]$ to diverge. The smoothed spectral absisca is therefore necessarily greater than $\alpha(\boldsymbol{W})$. Consequently, we may seek to minimize $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ instead of $\alpha(\boldsymbol{W})$, which is advantageous since $\tilde{\alpha}_\epsilon$ is a smooth function of the synaptic weights.

The tractability of the approach stems from the computability of $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ and its derivatives w.r.t $\boldsymbol{W}$. For any $s > \alpha(\boldsymbol{W})$, matrix $\boldsymbol{Q}(s)$ defined in Equation 3.11 is known to be the solution to the following Lyapunov equation

$$(\boldsymbol{W} - s\mathbb{1})^T \boldsymbol{Q}(s) + \boldsymbol{Q}(s) (\boldsymbol{W} - s\mathbb{1}) = -2 \cdot \mathbb{1} \tag{3.12}$$

Solving this equation numerically can be done efficiently (Bartels and Stewart, 1972). Knowing that $\mathrm{tr}\,[\boldsymbol{Q}(s)] - \epsilon^{-1}$ is a decreasing function of s, one can apply standard root-finding

methods to find $\tilde{\alpha}_\epsilon(\boldsymbol{W})$. Finally, Vanbiervliet et al. (2009) also worked out the gradient

$$\frac{\partial\, \tilde{\alpha}_\epsilon(\boldsymbol{W})}{\partial \boldsymbol{W}} = \frac{\boldsymbol{Q}(\tilde{\alpha}_\epsilon)\boldsymbol{P}(\tilde{\alpha}_\epsilon)}{\mathrm{tr}\,[\boldsymbol{Q}(\tilde{\alpha}_\epsilon)\boldsymbol{P}(\tilde{\alpha}_\epsilon)]} \tag{3.13}$$

where

$$\boldsymbol{P}(s) = 2\int_0^\infty e^{t(\boldsymbol{W}-s\mathbb{1})}\; e^{t(\boldsymbol{W}-s\mathbb{1})^T}\; dt \tag{3.14}$$

solves the Lyapunov equation dual to Equation 3.12:

$$(\boldsymbol{W} - s\mathbb{1})\,\boldsymbol{P}(s) + \boldsymbol{P}(s)\,(\boldsymbol{W} - s\mathbb{1})^T = -2\cdot\mathbb{1} \tag{3.15}$$

The iterative gradient descent on $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ entails the following steps:

1. Compute the current value of the smoothed spectral abscissa $\tilde{\alpha}_\epsilon(\boldsymbol{W})$. This implies multiple iterations of a numerical root-finding method (e.g. bisection) on $s \mapsto \mathrm{tr}[\boldsymbol{Q}(s)] - \epsilon^{-1}$. Each iteration requires solving Equation 3.12 for $\boldsymbol{Q}(s)$.

2. Solve Equation 3.12 and Equation 3.15 with $s = \tilde{\alpha}_\epsilon$ found in step 1. This gives matrices $\boldsymbol{Q}(\tilde{\alpha}_\epsilon)$ and $\boldsymbol{P}(\tilde{\alpha}_\epsilon)$, which must be multiplied to form the desired gradient (Equation 3.13).

3. Move the inhibitory weights by a small amount in the direction of the negative gradient. That is, for every existing inhibitory synapse w_{ij} (only 40% of all possible inhibitory connections exist at any given time, see step 6), set $w_{ij} \leftarrow w_{ij} - \eta\,(\partial\tilde{\alpha}_\epsilon/\partial\boldsymbol{W})_{ij}$. Here η is a learning rate.

4. Enforce constraint 1 (inhibition remains inhibition), by clipping positive inhibitory weights to zero

5. Enforce constraint 2 (global E/I balance) through Equation 3.6. This step is not necessary for stability optimization, but is essential to make sure that the high correlation of excitatory and inhibitory input currents that emerges from optimization is not overwhelmed by the baseline correlation contributed by shared population fluctuations (see main text and 3.5.2 in the Supplemental Data).

6. Enforce constraint 3 (connectivity sparsity). Remove any existing connection w_{ij} that step 4 may have set to zero, and replace it by another connection w_{ik} where inhibitory neuron k is chosen randomly. Set the strength of these new connections to zero initially. Again this constraint is not required, but adds to the biological plausibility of the resulting connectivity.

Steps 1 through 6 are then repeated until convergence of the spectral abscissa. In the Supplemental Data, we show how step 1 may be short-circuited, thereby significantly reducing the computational cost of the stabilization procedure (it also constrains the choice of ϵ, which so far has been left as free parameter). In any event, the above-described procedure with $\epsilon = 0.01$ does achieve stabilization and yields cISNs similar to the one presented in the main text (in particular, it builds ones with identical dynamical properties).

A key result presented here is that cISNs exhibit a detailed balance of E and I inputs during spontaneous activity (c.f. Figure 3.6). In particular, we make the point that this balance goes well beyond the "trivial" temporal correlations between E and I synaptic inputs that arise from the (necessarily shared) temporal fluctuations of the population-averaged firing rate. In order to make this point, we need to make sure that the population-averaged activity fluctuates as little as possible. We have shown previously (Hennequin et al., 2012) that this can be achieved by making inhibition γ times stronger than excitation on average, with $\gamma > 1$ (and this was shown in Renart et al. (2010) too, for networks of nonlinear threshold units). This explains why we choose $\gamma = 3$ in Equation 3.6, and also why we strive to maintain this inhibition dominance throughout the stabilization procedure (step 5 above).

3.4.5 Analysis of rotational dynamics

The planes of rotation of Figure 3.2H are found with jPCA, a dynamical variant of principal component analysis (Churchland et al., 2012). It is a method to extract planes of rotation from multidimensional time trajectories. Given data of the form $(\boldsymbol{y}(t), \dot{\boldsymbol{y}}(t)))$, where $\dot{\boldsymbol{y}}(t)$ denotes the temporal derivative of $\boldsymbol{y}(t)$, jPCA attempts to fit (through standard least-square regression) a linear oscillatory model of the form

$$\dot{\boldsymbol{y}}(t) = \boldsymbol{M}_{\text{skew}} \boldsymbol{y}(t) \tag{3.16}$$

where $\boldsymbol{M}_{\text{skew}}$ is a skew-symmetric matrix, therefore one with purely imaginary eigenvalues. The two leading eigenvectors of the best-fitting $\boldsymbol{M}_{\text{skew}}$ (associated with the largest conjugate pair of imaginary eigenvalues) define the plane in which the trajectory rotates most strongly. Here we compute the jPCA projections exactly as prescribed in Churchland et al. (2012), using the gradient implementation we derive in subsection 3.5.3. The data consist of the population responses during the first 200ms following the go cue for each of our 27 conditions, sampled in 1ms time steps. To make sure that the jPC projection captures enough of the data variance, that is, that the observed rotational dynamics (if any) are significant, the data is first projected down to its top 6 standard principal components.

3.4.6 Muscle activation through linear readouts

In Figure 3.4, a single pair of muscle readouts is learned from 200 training trials (100 trials for each of the "snake" and "butterfly" movements). We assume the following linear model:

$$\boldsymbol{z}_t = (\boldsymbol{m}_1; \boldsymbol{m}_2)^T \, g\left[\boldsymbol{x}_t\right] + \boldsymbol{b} + \boldsymbol{\varepsilon}_t \tag{3.17}$$

where $\boldsymbol{z}_t$ (size 2) denotes the target muscle activation vector at discrete time t, $g\left[\boldsymbol{x}_t\right]$ is the $N \times 1$ vector of momentary firing rates in the network, and $\boldsymbol{\varepsilon}_t$ is the vector of residual errors (size 2). The readout weights (column vectors $\boldsymbol{m}_i$, $i = 1, 2$) are parameters which we optimize through simple least-square regression, together with a pair of biases ($\boldsymbol{b}$). The snake (resp. butterfly) target trajectory is made of 58 points (resp. 26 points), equally spaced in time over 500ms following the go cue, which defines the discrete time variable t in Equation 3.17. The activity vector $\boldsymbol{x}_t$ is sampled accordingly for each movement.

3.5 Supplemental Data

3.5.1 Optimal stabilization of recurrent networks

In this section, we argue that step 1 in the optimization procedure given in the Methods can be bypassed altogether, which significantly reduces the computational cost, and turns out to promote faster stabilization. Indeed, in the main text we have left ϵ a free parameter. ϵ modulates the distance between the spectral abscissa α and its upper bound $\tilde{\alpha}_\epsilon$: if ϵ decreases, $\tilde{\alpha}_\epsilon$ becomes a tighter upper bound to the spectral abscissa. In pilot studies, we realized that stability could be reached much faster if ϵ was set to decrease progressively during the course of the gradient descent. Empirically, it seemed a good idea to keep the ratio $\tilde{\alpha}_\epsilon/\alpha$ constant, and to adjust ϵ in every iteration to meet this need. Maintaining such a ratio constant resulted in an exponential decay of the spectral abscissa during the course of optimization. Mathematically, this means that the cost function $(\boldsymbol{W} \mapsto \tilde{\alpha}_\epsilon(\boldsymbol{W}))$ keeps moving, but it becomes a progressively tighter upper bound on α, and brings a crucial advantage: one no longer needs to compute $\tilde{\alpha}_\epsilon$!

We thus capitalize on this observation and set $\tilde{\alpha}_\epsilon(\boldsymbol{W}) = C\alpha(\boldsymbol{W})$ in every iteration, with $C = 1.5$ (empirically good choice). Note that this automatically constrains ϵ to a value of $1/\mathrm{tr}[\boldsymbol{Q}(C\alpha)]$, where $\boldsymbol{Q}(\cdot)$ is defined in Equation 3.11 in the Methods. Steps 2 to 6 are then performed as prescribed in the Methods. Note that computationally, the cost is still of order $K \cdot N^3$, but the large constant K implied by the iterative root-finding method of Step 1 is dramatically reduced. Note also that Step 2 requires solving Equation 3.12 and Equation 3.15, for which a single Schur decomposition of $\boldsymbol{W}$ needs to be computed. As a byproduct, the Schur decomposition also returns the spectral abscissa at no further cost, so α needs not be computed separately.

The above simplified procedure is very effective, up to one small detail. Since $\alpha(\boldsymbol{W})$ is non-smooth, one would like to keep the smoothed spectral abscissa some safe margin away from α, so its gradient remains well-behaved. In the above scenario, $C\alpha$ becomes increasingly closer to α as stability optimization progresses. This indeed leads to unstable learning as α becomes as low as ~ 0.2. In every iteration, we therefore set the smoothed spectral abscissa to $C\alpha$ or $\alpha + B$, whichever was the greatest. We use $B = 0.2$.

As mentioned above, if ϵ is set to decrease progressively during the course of optimization, the whole stabilization procedure looses its interpretation as a gradient descent on a fixed objective function. Nevertheless, the (moving) cost function remains invariably an upper bound on the spectral abscissa (which becomes tighter and tighter), so there is no need to worry about this so long as mere stabilization is the only real objective.

3.5.2 How much do shared population fluctuations contribute to the detailed E/I balance?

Intuitively, fluctuations of the mean population firing rate, by definition shared by all neurons, are expected to increase the match between the E and I input currents into single cells. To make this intuition precise and to dissect the impact of population-wide fluctuations on the detailed balance, we study the following simplified model.

We consider a random balanced network of size $N = 2M \gg 1$, in which excitatory weights (including the zero weights) are drawn independently from some distribution with mean $\mu_E/\sqrt{N}$ and variance σ_E^2/N, and similarly for the inhibitory weights with parameters $-\mu_I/\sqrt{N}$ and σ_I^2/N. The key simplification is to forget about the recurrent dynamics, and to assume that, due to shared population-wide fluctuations, the neuronal activities $x_i(t)$ comprise some common fluctuation $r(t)$ and private fluctuations $\xi_i(t)$, in the following mixing proportions:

$$x_i(t) = r(t)\sqrt{c} + \xi_i(t)\sqrt{1-c} \tag{3.18}$$

Here $r(t)$ and the $\xi_i(t)$ are all stationary fluctuations of unit variance, without loss of generality. Note that the temporal aspects are irrelevant to this discussion – only stationarity matters. Clearly this simplification neglects the width of the distribution of pairwise correlations, and focuses on its mean c. We interpret the first half of the x_i's ($1 \le i \le M$) as the activity of the excitatory neurons, and the second half as the activity of the inhibitory neurons. By design, we have

$$\langle x_i(t)x_j(t)\rangle = (1-c)\delta_{ij} + c \tag{3.19}$$

where $\langle\cdot\rangle$ denotes temporal averaging. We now form the E and I input currents "artificially" by passing $\boldsymbol{x}(t)$ through the weight matrix $\boldsymbol{W}$. We may thus write the E and I input currents to neuron ℓ as

$$c_\ell^E(t) = \sum_{j=1}^{M} w_{\ell j}x_j(t) \qquad \text{and} \qquad c_\ell^I(t) = -\sum_{j=M+1}^{N} w_{\ell j}x_j(t) \tag{3.20}$$

The E/I input correlation for cell ℓ is defined as

$$\rho_\ell \quad = \quad \frac{\langle c_\ell^E(t)\cdot c_\ell^I(t)\rangle}{\sqrt{\langle c_\ell^E(t)^2\rangle \cdot \langle c_\ell^I(t)^2\rangle}} \tag{3.21}$$

We then compute

$$\langle c_\ell^E(t)\cdot c_\ell^I(t)\rangle \quad = \quad -\sum_{j=1}^{M}\sum_{k=M+1}^{N} w_{\ell j}w_{\ell k}\langle x_j(t)x_k(t)\rangle \tag{3.22}$$

$$= \quad -c\sum_{j=1}^{M}\sum_{k=M+1}^{N} w_{\ell j}w_{\ell k} \tag{3.23}$$

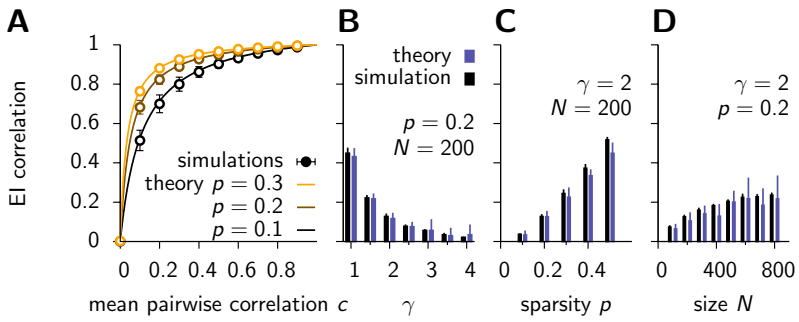


Figure 3.7: **How does shared neural variability influence the detailed E/I balance?** (**A**) Pearson correlation coefficient between E and I input currents into single cells, averaged across a population of size $N = 200$, as a function of the mean pairwise correlation between neuronal firing rates. Errorbars denote one standard deviation across all neurons. The solid curves represent Equation 3.28. The network dynamics were not simulated as such: instead, individual neuronal activities were generated independently with some shared baseline (see text, Equation 3.18). Input currents were "artifically" computed by passing the network activity vector through the E and I synaptic weights, which were drawn as indicated in Equation 3.27. (**B**,**C** and **D**) The full network dynamics were simulated, and we report here the mean E/I correlation coefficient obtained from the simulation (black), or the coefficient obtained from the theory (Equation 3.28) in which the average pairwise correlation coefficient c was nonetheless numerically estimated from the network dynamics (blue). Errorbars denote one standard deviation across 10 trials (each with a new connectivity matrix $\boldsymbol{W}$ drawn following Equation 3.27).

Since the $w_{\ell j}$ and $w_{\ell k}$ have been drawn independently from their respective distributions, we may approximate the double sum by M^2 times the product of their means, i.e. $-\sum_j \sum_k w_{\ell j} w_{\ell k} \simeq M^2 \mu_E \mu_I / N$, leading to

$$\langle c_\ell^E(t) \cdot c_\ell^I(t) \rangle = \frac{N \mu_E \mu_I c}{4} \tag{3.24}$$

A very similar calculation yields

$$\langle c_\ell^E(t)^2 \rangle = \frac{cN\mu_E^2}{4} + \frac{(1-c)}{2}\left(\mu_E^2 + \sigma_E^2\right) \tag{3.25}$$

and similarly for $\langle c_\ell^I(t)^2 \rangle$. We thus conclude

$$\rho_\ell \quad = \quad \frac{1}{\sqrt{\left[1 + \frac{2(1-c)}{Nc}\left(1 + \frac{\sigma_E^2}{\mu_E^2}\right)\right]\left[1 + \frac{2(1-c)}{Nc}\left(1 + \frac{\sigma_I^2}{\mu_I^2}\right)\right]}} \tag{3.26}$$

which no longer depends on cell index ℓ, and converges to 1 as $N \to \infty$ (if c is non-zero). At this point we may conclude that, if population fluctuations are finite (non-zero), then a perfect balance is established between E and I currents into single cells, so long as N is large enough. However, when inhibition dominates over excitation ($\mu_I > \mu_E$), the mean pairwise correlations coefficient c vanishes with large N (Renart et al., 2010; Hennequin et al., 2012); in fact, it scales as $1/N$. We therefore expect ρ_{EI} to have a finite limit as $N \to \infty$.

To illustrate the result of Equation 3.26, we generated random sparse balanced networks, with connectivity matrix $\boldsymbol{W}$ drawn as follows:

$$w_{ij} = \frac{1}{\sqrt{N}} \cdot \left\{ \begin{array}{ll} +w_0 & \text{if } j \leq N/2 \\ -\gamma w_0 & \text{if } j > N/2 \end{array} \right] \text{ with proba. } p \\ 0 \text{ with proba. } (1-p) \tag{3.27}$$

Simple algebra yields the respective means and variances of the E and I weights, which we plug in Equation 3.26 to receive

$$\rho = \frac{1}{1 + \frac{2(1-c)}{Npc}} \tag{3.28}$$

This result is plotted in Figure 3.7A, together with numerical simulations of this simplified problem (where the $x_i(t)$ are generated by Equation 3.18).

In order to further validate Equation 3.26 in situations were neuronal activities are actually generated by the recurrent dynamics, we simulated spontaneous activity as described in the Experimental Procedures. To vary the mean pairwise correlation coefficient c, with varied γ, p, and N independently. The results are shown in Figure 3.7B-D. In particular, Figure 3.7D confirms that ρ converges to a value smaller than 1 as the network grows in size. This is because, when $\gamma > 1$, the variance of the population fluctuations scales with $1/N$ (Hennequin et al., 2012).

3.5.3 Derivation of gradient-based jPCA

jPCA is a dynamical variant of principal component analysis (PCA) that seeks to discover planes in which a multidimensional time trajectory rotates most strongly. If there is a rotational component to the collective dynamics of a set of units, it will show up in the top jPC planes. The original description of the method is provided in the Supplementary Information of Churchland et al. (2012), and is an excellent reference. There it is mentioned in passing that for large datasets, it may be advantageous to perform the least-square optimization that jPCA entails using gradient methods. In this section I provide a (straightforward) derivation of the gradient in question, which then can be used to solve the jPCA problem using any numerical gradient-based optimization routine.

The data on which jPCA operates is a collection of K snapshots of network activity and associated instantaneous time-derivatives : $\left\{\left(x^k, \dot{x}^k\right)\right\}_{k=1\ldots K}$. jPCA seeks the best possible rotational linear dynamics description of the data, i.e. an $N \times N$ skew-symmetric matrix M_{skew} such that the error made by the prediction model

$$\dot{x} = M_{\text{skew}} x \tag{3.29}$$

is minimized. The error is defined in the standard least-square sense:

$$\mathcal{L}(M_{\text{skew}}) \quad = \quad \frac{1}{2}\sum_{k=1}^{K}\left\|\dot{x}^k - M_{\text{skew}}x^k\right\|^2 \tag{3.30}$$

In a skew-symmetric matrix, there are only $N(N-1)/2$ degrees of freedom. One therefore looks for strict lower triangles T, and one goes from T to a full skew-symmetric matrix via a linear map $H(T) = T - T^T$, where $\cdot^T$ denotes the transpose. The error to be minimized is therefore

$$\mathcal{L}(T) \quad = \quad \frac{1}{2}\sum_{k=1}^{K}\left\|\dot{x}^k - \left(T - T^T\right)x^k\right\|^2 \tag{3.31}$$

Let us express the model-reconstructed momentary derivative, which we call $\overset{\star}{x}^k$:

$$\overset{\star}{x}{}^k_m \quad = \quad \left[\left(T - T^T\right)x^k\right]_m \quad = \quad \sum_{n=1}^{m-1} T_{m,n}x^k_n - \sum_{n=m+1}^{N} T_{n,m}x^k_n \tag{3.32}$$

Thus for $a > b$

$$\frac{\partial\, \overset{\star}{x}{}^k_m}{\partial T_{a,b}} \quad = \quad \delta_{a.m}\cdot x^k_b - \delta_{m,b}\cdot x^k_a \tag{3.33}$$

Let us differentiate the error in Equation 3.31 w.r.t $T_{a,b}$, with $a > b$:

$$\frac{\partial\mathcal{L}}{\partial T_{a,b}} \quad = \quad \frac{1}{2}\sum_{k=1}^{K}\sum_{m=1}^{N}\left(\overset{\star}{x}{}^k_m - \dot{x}^k_m\right)^2 \tag{3.34}$$

$$= \quad -\sum_{k=1}^{K}\sum_{m=1}^{N}\left(\delta_{a,m}x^k_b - \delta_{b,m}x^k_a\right)\left(\overset{\star}{x}{}^k_m - \dot{x}^k_m\right) \tag{3.35}$$

$$= \quad -\sum_{k=1}^{K}\left[x^k_b\cdot\left(\overset{\star}{x}{}^k_a - \dot{x}^k_a\right) - x^k_a\cdot\left(\overset{\star}{x}{}^k_b - \dot{x}^k_b\right)\right] \tag{3.36}$$

The final expression is the gradient one should use to perform batch optimization and obtain the least-squares solution for T. The objective function is convex, so gradient methods are guaranteed to converge to the global minimum.

3.5.4 Supplementary Figures

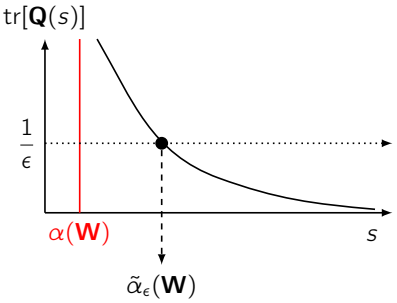


Figure 3.8: **Schematics of the smoothed spectral abscissa**. The decreasing and convex behavior of tr[$\boldsymbol{Q}(s)$] as a function of s is sketched in black. It goes to zero as $s \to +\infty$, and diverges as s approaches the spectral abscissa $\alpha(\boldsymbol{W})$ from above. The point on the x-axis at which the curve crosses the dotted line at $1/\epsilon$ defines the smoothed spectral abscissa $\tilde{\alpha}_\epsilon(\boldsymbol{W})$.

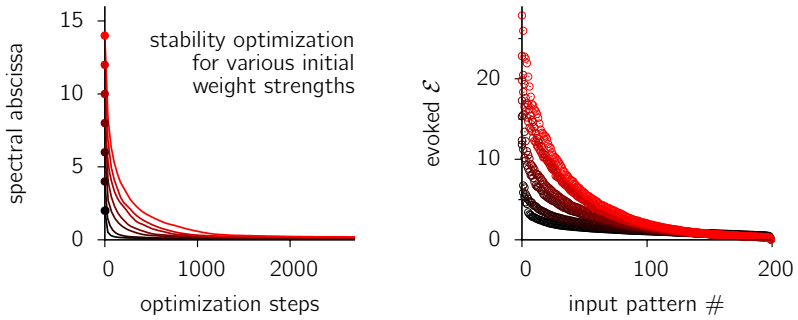


Figure 3.9: **Optimizing random networks with various initial weight strengths**. (**Left**) Stability is achieved no matter how strong the excitatory synaptic weights are. The curves are obtained using the exact same procedure as we used to generate the network of Figure 3.2, starting from random balanced networks with different initial spectral abscissae (color-coded). Note that in this model of random balanced connectivity, the spectral abscissa uniquely determines the strength of the E synaptic weights. (**Right**) Corresponding "energy profiles" after optimization (see caption for Figure 3.3A). The expected energy (average of the energy curve) increases with the average E weight strength.

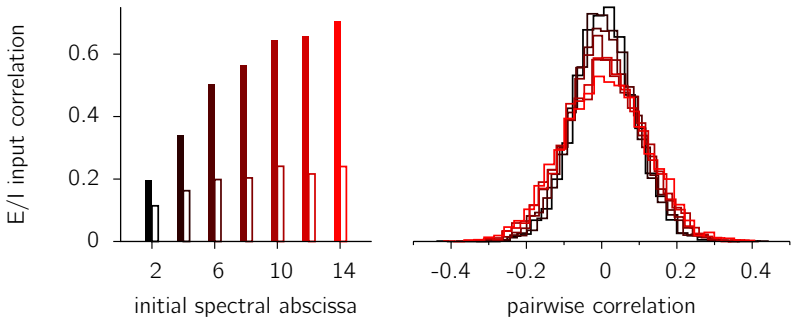


Figure 3.10: **The stronger the weights, the finer the E/I balance in stabilized networks, and the wider the distribution of pairwise correlations**. The color code is the same as in Figure 3.9. (**Left**) Empty boxes denote the average correlation coefficient of E and I inputs when they are taken in *different* cells; solid boxes denote the average correlation of E and I inputs into the same cell. (**Right**) Distributions of pairwise correlations during spontaneous activity.

CHAPTER 4

Towards an inhibitory synaptic plasticity rule for optimal and robust network stabilization

In chapter 3, I have mostly focused on the dynamical behavior of networks in which recurrent excitation is complex and strongly unstable on its own, but stabilized by fine-tuned inhibitory feedback. To instantiate such networks, I have used an optimal stabilization procedure derived from control-theoretic principles. I have entirely left aside the question of *how* inhibitory stabilization could be achieved in the cortex, for example through realistic synaptic plasticity mechanisms. The present (small) chapter is intended to revisit this question. A large part is dedicated to understanding the nature of the inhibitory feedback optimization performed in chapter 3, by discussing some important properties of the "smoothed spectral abscissa". I then establish a link between the minimization of the smoothed spectral abscissa (the measure of robust stability introduced in chapter 3, see also Vanbiervliet et al. (2009)) and the phenomenological model of inhibitory synaptic plasticity recently developed in Vogels et al. (2011) to account for the self-organization of a precise excitation/inhibition balance along the auditory synaptic pathway of the rat. Overall, this chapter shows that *spontaneous circuit dynamics* can be used as a substrate for robust network stabilization through inhibitory synaptic plasticity.

I begin by formulating the assumptions made for the network dynamics, and I state the problem of "robust" network stabilization. I then explain why traditional spectral analysis is ill-suited to the problem of robust stabilization, and introduce again the smoothed spec-

tral abscissa (chapter 3) as a better alternative. The latter can be used to formulate a parameter-free version of the robust stabilization problem. This formulation turns out to be stricly equivalent to the minimization of the average variance of the single-neuron *spontaneous* activity fluctuations in the linear regime. Interestingly, such minimization can be approximated by a local learning rule at inhibitory synapses that is structurally similar to the plasticity rule of Vogels et al. (2011), itself inspired by the experimental results of Woodin et al. (2003). I show numerically that the local learning rule can indeed be used to stabilize complex (nonlinear) networks based on their spontaneous activity. As a local approximation to a global problem, it is necessarily suboptimal, but yields solutions (inhibitory synaptic strengths) that correlate with the solutions found by the optimal control-theoretic method.

4.1 Spontaneous rate dynamics and network stability

4.1.1 Setup

Here I briefly recall the formalism I have used previously (chapter 2 and chapter 3) to describe balanced cortical dynamics on the level of firing rates. I consider a network of N neurons recurrently connected to one another through a matrix $\boldsymbol{W}$ of synaptic interactions. I assume that $\boldsymbol{W} \equiv \boldsymbol{W}(\boldsymbol{z})$ depends on a certain number of parameters, summarized in a vector $\boldsymbol{z}$, and that only those parameters may be optimized to stabilize the network (see below). For example, in chapter 3, $\boldsymbol{z}$ was simply made of all the inhibitory synaptic efficacies.

The dynamics of each neuron i are described by a single activation variable $x_i(t)$, from which a firing rate $g[x_i(t)]$ is computed where $g[\cdot]$ is the input-output nonlinearity (reminiscent of the neuronal f-I curve). Following initialization in some non-zero activity state $\boldsymbol{x}_0$, the noiseless free dynamics of the network are described by

$$\tau \frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{W} g[\boldsymbol{x}(t)] + \delta(t)\gamma \frac{\boldsymbol{x}_0}{||x_0||} \tag{4.1}$$

where $\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))^T$ is the column vector of activation variables, $\delta(t)$ denotes the Dirac delta function, and γ sets the overall magnitude of the initial condition. I also define spontaneous activity as the recurrent processing of unspecific noisy external inputs, following

$$\tau \frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{W} g[\boldsymbol{x}(t)] + \gamma \boldsymbol{\xi}(t) \tag{4.2}$$

where $\boldsymbol{\xi}(t)$ is a spatially and temporally white N-dimensional Wiener process of unit variance and γ again sets the overall magnitude of that input noise.

4.1.2 Evoked energy and amplification factor

Two definitions will be useful for the rest of this chapter. First, I define the average "evoked energy" $\mathcal{E}_0(\boldsymbol{W})$ in the noiseless free dynamics of Equation 4.1 as

$$\mathcal{E}_0(\boldsymbol{W}) \quad \stackrel{\text{def}}{=} \quad \left\langle \int_0^\infty \|g\left[\boldsymbol{x}(t)\right]\|^2 \, dt \right\rangle_{\boldsymbol{x}_0 \rightsquigarrow \mathcal{N}(0,\mathbb{1})} \tag{4.3}$$

where $\langle \cdot \rangle_{\boldsymbol{x}_0 \rightsquigarrow \mathcal{N}(0,\mathbb{1})}$ denotes the expectation over initial condition $\boldsymbol{x}_0$, drawn from a multivariate Gaussian distribution with identity covariance matrix. Second, I define the "amplification factor" $A(\boldsymbol{W})$ as the average variance of the single-cell spontaneous activity fluctuations in the context of Equation 4.2:

$$A(\boldsymbol{W}) \quad \stackrel{\text{def}}{=} \quad \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \tag{4.4}$$

Here σ_i^2 denotes the temporal variance of the firing rate $g\left[x_i(t)\right]$ of unit i during spontaneous activity.

From now on I will restrict the analysis to *linear* networks, i.e. to situations where γ is small enough that, given a stable $\boldsymbol{W}$, firing rates deviate only weakly around baseline. For simplicity and without loss of generality I now set $\gamma = \sqrt{2/\tau}$, so that both $\mathcal{E}_0(\boldsymbol{W})$ and $A(\boldsymbol{W})$ equal 1 for an unconnected network ($\boldsymbol{W} = 0$).

In the linear regime, the dynamics of Equation 4.2 become those of a multivariate Ornstein-Uhlenbeck process (as in chapter 2). In this case, the covariance matrix $\left\langle \boldsymbol{x}(t)\boldsymbol{x}(t)^\dagger \right\rangle_t$ of the network activity is given by $\boldsymbol{P}(\boldsymbol{W}, 1)$ (Gardiner, 1985), defined more generally as

$$\boldsymbol{P}(\boldsymbol{W}, s) \quad \stackrel{\text{def}}{=} \quad \int_0^\infty \left(e^{t(\boldsymbol{W} - s\cdot\mathbb{1})}\right) \left(e^{t(\boldsymbol{W} - s\cdot\mathbb{1})}\right)^T dt \tag{4.5}$$

(the more general form $\boldsymbol{P}(\boldsymbol{W}, s)$ will be useful later). The sum in Equation 4.4 can be written as the trace of the covariance matrix, such that

$$A(\boldsymbol{W}) \quad = \quad \frac{\mathrm{tr}\left[\boldsymbol{P}(\boldsymbol{W}, 1)\right]}{N} \tag{4.6}$$

We have seen in chapter 3 that the average evoked energy $\mathcal{E}_0(\boldsymbol{W})$ can be written in a similar form as

$$\mathcal{E}_0(\boldsymbol{W}) \quad = \quad \frac{\mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, 1)\right]}{N} \tag{4.7}$$

where again $\boldsymbol{Q}(\boldsymbol{W}, 1)$ is defined more generally as

$$\boldsymbol{Q}(\boldsymbol{W}, s) \quad \stackrel{\text{def}}{=} \quad \int_0^\infty \left(e^{t(\boldsymbol{W} - s\cdot\mathbb{1})}\right)^T \left(e^{t(\boldsymbol{W} - s\cdot\mathbb{1})}\right) dt \tag{4.8}$$

Using the identity $\mathrm{tr}\,(\boldsymbol{AB}) = \mathrm{tr}\,(\boldsymbol{BA})$ together with Equations 4.5 and 4.8, we note that in the linear regime, the amplification factor and the expected evoked energy are one and the same:

$$\mathcal{E}_0(\boldsymbol{W}) \quad = \quad A(\boldsymbol{W}) \tag{4.9}$$

4.1.3 Linear stability and associated caveats

As in the previous chapters, the stability of the network dynamics is understood here in the linear sense. That is, the network is said to be stable if and only if the linear versions of the dynamics in equations 4.1 and 4.2 lead to finite $\mathcal{E}_0(\boldsymbol{W})$ (or finite $A(\boldsymbol{W})$). Linear stability depends solely on the matrix of synaptic interactions $\boldsymbol{W}$. More specifically, it depends on its eigenvalue spectrum: for the network to be stable, all the eigenvalues of $\boldsymbol{W}$ must have real part smaller than one. Therefore, the relevant quantity for stability is the spectral abscissa $\alpha(\boldsymbol{W})$ defined as

$$\alpha(\boldsymbol{W}) \quad \stackrel{\text{def}}{=} \quad \max_{\text{eigv. } \lambda} \mathrm{Re}(\lambda) \tag{4.10}$$

Importantly, linear stability is only about the *asymptotic* behavior of the network following an arbitrary initial condition: when $\alpha(\boldsymbol{W}) < 1$, after enough time has passed, the firing rates decay back to rest exponentially fast. Indeed, one can always find a constant K such that, for sufficiently large t, $\|\boldsymbol{x}(t)\| \leq K \exp\left[(\alpha(\boldsymbol{W}) - 1)t/\tau\right]$. This ensures that the expected evoked energy $\mathcal{E}_0(\boldsymbol{W})$ remains finite. The *transient* behavior of the network, however, is not captured by standard eigenvalue analysis (Trefethen and Embree, 2005). For a large class of matrices known as "non-normal" (c.f. chapter 2), stability can coexist with transient amplification of large amplitude that is due to purely linear effects[1]. Thus, although the expected evoked energy $\mathcal{E}_0(\boldsymbol{W})$ (or amplification $A(\boldsymbol{W})$ in the spontaneous case) is finite, it may still be large enough to cause the system to leave the linear regime, and to perhaps undergo an unwanted transition into unstable dynamics (we have seen an example of such a transition in Figure 3.3 of chapter 3). Ideally, one would like to have a more robust stability certificate than $\alpha(\boldsymbol{W}) < 1$, i.e. one that takes into account such transient amplification effects. The smoothed spectral abscissa introduced in chapter 3 provides precisely this[2], as we shall see below.

In the brain, synaptic connections undergo constant changes, on many levels ranging from transient changes in the effective connectivity due to synaptic transmission failure and short-

[1] In chapter 2, transient amplification was revealed by a Schur decomposition of $\boldsymbol{W}$, which is nothing but linear algebra.

[2] There is also an extensive theory of "pseudospectra", a concept developed in the early 1990s primarily to fill the gaps left by standard spectral analysis with respect to transient behavior; I warmly recommand Trefethen and Embree (2005) for a comprehensive overview of the mathematics and physics literature on these issues.

term depression/facilitation, to more enduring patterns such as neuronal death, spine addition/removal, and synaptic plasticity. Potentially, some of those changes may threaten stability. The spectral abscissa criterion for stability does not take this into account: it only says whether or not the network is stable in its current state, but does not quantify how much damage the connectivity is allowed to suffer before the network runs unstable. As we shall see below, the smoothed spectral abscissa is a more satisfying stability measure in this respect too.

More pragmatically, there is a third reason why the spectral abscissa is not necessarily a good cost function for network stabilization: it is not smooth in the synaptic weights, hence it is difficult to minimize (Burke et al., 2002; Noll and Apkarian, 2005). In contrast, as the name indicates, the smoothed spectral abscissa is differentiable everywhere, allowing the use of gradient methods as I have shown in chapter 3.

4.2 Smooth and robust formulation of the stabilization problem

4.2.1 The ϵ-smoothed spectral abscissa

Intuitively, enough negative self-feedback in each unit should ensure that, despite strong interactions in $\boldsymbol{W}$, the firing rates do not blow up. By "negative self-feedback", I mean replacing each synaptic weight w_{ii} by $w_{ii} - s$ with $s > 0$, which in matrix notation reads $\boldsymbol{W} \leftarrow \boldsymbol{W} - s\mathbb{1}$. The spectral abscissa $\alpha(\boldsymbol{W})$ can in fact be defined as the minimal value that s would need to take to guarantee stability[3], that is, to guarantee that $\mathcal{E}_0(\boldsymbol{W} - s\mathbb{1})$ is finite. Let us rephrase this using Equation 4.7:

$$\alpha(\boldsymbol{W}) \quad = \quad \inf\left\{s \in \mathbb{R} \;/\; \mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, s)\right] < \infty\right\} \tag{4.11}$$

The $< \infty$ criterion makes it clear that $\alpha(\boldsymbol{W})$ is an asymptotic quantity (the infimum is not within the set inside curly brackets). A more robust stability measure is obtained by relaxing this inequality, and requiring $\mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, s)\right] \leq N/\epsilon$ for some small positive ϵ instead. This is precisely[4] how Vanbiervliet et al. (2009) defined the ϵ-smoothed spectral abscissa $\tilde{\alpha}_\epsilon(\boldsymbol{W})$:

$$\tilde{\alpha}_\epsilon(\boldsymbol{W}) \quad \stackrel{\text{def}}{=} \quad \inf\left\{s \in \mathbb{R} \;/\; \mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, s)\right] \leq \frac{N}{\epsilon}\right\} \tag{4.12}$$

[3] This is because the spectrum of $\boldsymbol{W} - s\mathbb{1}$ is nothing but the spectrum of $\boldsymbol{W}$ shifted by s units to the left, so that in particular $\alpha(\boldsymbol{W} - s\mathbb{1}) = \alpha(\boldsymbol{W}) - s$.

[4] Vanbiervliet et al. used $\leq 1/\epsilon$ instead of $\leq N/\epsilon$ for the relaxation. We make this slight modification here because, in the context of balanced neuronal network in which the synaptic weights are of order $1/\sqrt{N}$, the evoked energy remains finite in the large N limit (c.f. chapter 2). That is, $\mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, 1)\right] = \mathcal{O}(N)$.

The function $s \mapsto \mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, s)\right]$ can be shown to be monotonically decreasing (and convex), so $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ is simply the unique solution to

$$\mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, \tilde{\alpha}_\epsilon)\right] = \frac{N}{\epsilon} \tag{4.13}$$

This definition was already illustrated in Figure 3.8 in chapter 3. It is easily shown that $\tilde{\alpha}_\epsilon$ is a growing function of ϵ. In the limit of very small ϵ, the definition of $\tilde{\alpha}_\epsilon$ collapses to that of α, and indeed $\lim_{\epsilon \to 0^+} \tilde{\alpha}_\epsilon = \alpha$ (Vanbiervliet et al., 2009).

Importantly, the smoothed spectral abscissa bounds the spectral abscissa from above:

$$\alpha(\boldsymbol{W}) \quad < \quad \tilde{\alpha}_\epsilon(\boldsymbol{W}) \qquad (\forall \epsilon > 0) \tag{4.14}$$

This implies that the condition

$$\tilde{\alpha}_\epsilon(\boldsymbol{W}) \quad \leq \quad 1 \tag{4.15}$$

can be used as a sufficient condition for network stability, as it implies $\alpha(\boldsymbol{W}) < 1$. Moreover, it is a condition for robust stability. Indeed, when $\tilde{\alpha}_\epsilon(\boldsymbol{W}) = 1$, not only do we know the network is stable, we also know that $\mathcal{E}_0(\boldsymbol{W}) = 1/\epsilon$, which is easily checked by setting $\tilde{\alpha}_\epsilon(\boldsymbol{W}) = 1$ in Equation 4.13 and looking at Equation 4.7. More generally, since $\mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, s)\right]$ is a decreasing function of s, we have:

$$\tilde{\alpha}_\epsilon(\boldsymbol{W}) \leq 1 \quad \Rightarrow \quad \mathcal{E}_0(\boldsymbol{W}) \leq \frac{1}{\epsilon} \tag{4.16}$$

(and by Equation 4.9, the same holds for $A(\boldsymbol{W})$). Thus, if condition 4.15 holds for sufficiently large ϵ, we may obtain a reasonably useful upper bound on $\mathcal{E}_0(\boldsymbol{W})$, which may for example tell us that random disturbances in the network dynamics (e.g. external noise) are not to be amplified enough to push the system in the nonlinear regime and (potentially) distabilize the dynamics. Note, however, that the bound in Equation 4.16 applies to the *expected* evoked energy $\mathcal{E}_0(\boldsymbol{W}) \leq 1/\epsilon$ when the initial condition is chosen randomly without any knowledge of $\boldsymbol{W}$. Recalling that $\mathcal{E}_0(\boldsymbol{W}) = \mathrm{tr}\left[\boldsymbol{Q}(\boldsymbol{W}, 1)\right]/N$, it is clear that even if the trace[5] of $\boldsymbol{Q}$ is equal to $1/\epsilon$, $\boldsymbol{Q}$ could well have one single eigenvalue equal to N/ϵ, and $N-1$ negligible eigenvalues of order $1/N\epsilon$. In this case, the recurrent network would be ultra sensitive to patterns of inputs aligned onto the leading eigenvector of $\boldsymbol{Q}(\boldsymbol{W}, 1)$. In other words, "robust stability" as assessed by condition 4.15 is about the robustness to random "uninformed" input perturbations, and in principle does not conflict with selective input amplification – which we saw in chapter 3 can be functionally relevant.

The stability criterion of Equation 4.15 based on $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ also leads to a nice guarantee on the "distance to instability" of the synaptic connectivity. Indeed, Vanbiervliet et al. (2009)

[5] Recall that the trace of a matrix is also the sum of its eigenvalues – here $\boldsymbol{Q}(\boldsymbol{W}, 1)$ is a symmetric, positive semi-definite matrix, so all its eigenvalues are real and positive.

also showed that

$$\tilde{\alpha}_\epsilon(\boldsymbol{W}) \leq 1 \quad \Rightarrow \quad \beta(\boldsymbol{W}) \leq \frac{\epsilon}{2} \tag{4.17}$$

where

$$\beta(\boldsymbol{W}) \quad \stackrel{\text{def}}{=} \quad \min \left\{ \|\boldsymbol{D}\| \ : \ \boldsymbol{D} \in \mathbb{C}^N, \alpha(\boldsymbol{W} + \boldsymbol{D}) \geq 1 \right\} \tag{4.18}$$

is the minimum size of an additive matrix perturbation that would distabilize the network. Equation 4.17 is to be interpreted as follows: if for some $\epsilon > 0$, $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ can be made smaller than 1 by optimizing over some parameters (such as the inhibitory synaptic strengths), then the network is robustly stable in the sense that random additive disturbances $\boldsymbol{D}$ in the synaptic weights would need to be of "size" greater than $\epsilon/2$ to make the network unstable again. Here "size" is understood as the operator norm $\|\boldsymbol{D}\| = \sup_{\|z\|=1} \|\boldsymbol{D}z\|$, which is also the largest singular value of $\boldsymbol{D}$. Assuming that the noisy synaptic perturbations d_{ij} are independent of one another, and all drawn from the same distribution with zero mean and variance σ^2, then the singular values of $\boldsymbol{D}$ are scattered in the range $\left[0 : 2\sigma\sqrt{N}\right]$ following a "quarter circle" distribution (Mehta, 2004), so $\|\boldsymbol{D}\| \simeq 2\sigma\sqrt{N}$. Therefore, one may conclude that, so long as the standard deviation σ of the perturbations is smaller than $\epsilon/4\sqrt{N}$, the network remains stable.

Finally, the fact that $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ (unlike α) is a smooth function[6] of $\boldsymbol{W}$, and because the corresponding derivatives can be computed efficiently, $\tilde{\alpha}_\epsilon$ is a convenient cost function for network stabilization. As I have shown in chapter 3, the parameters $\boldsymbol{z}$ upon which $\boldsymbol{W}$ depends may be iteratively refined following the negative gradient of $\tilde{\alpha}_\epsilon$, until a sufficiently small value of α is reached.

4.2.2 Parameter-free robust stabilization and link to the Vogels rule

In the above discussion, the choice of $\epsilon > 0$ has been left arbitrary. In chapter 3 also, ϵ was chosen empirically, and set to progressively decrease during the course of optimization to obtain faster convergence. The previous section has made the point that, in the lucky event that the minimization of $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ does yield a solution smaller than one (e.g through constrained gradient descent as in chapter 3), the obtained network stability is augmented with the following two properties

- random inputs, on average, are not amplified by more than $1/\epsilon$ (in terms of evoked energy, Equation 4.3)

[6] This is because $\mathrm{tr}\,[\boldsymbol{P}(\boldsymbol{W}, s)]$ itself is a smooth function of $\boldsymbol{W}$ and s, and because $\tilde{\alpha}_\epsilon$ is defined implicitly as the unique solution of $\mathrm{tr}\,[\boldsymbol{P}(\boldsymbol{W}, s)] = 1/\epsilon$.

- (zero-mean) random damage in the synaptic weights can be as large as $\epsilon/4\sqrt{N}$ in standard deviation without threatening stability.

These two properties constitute what I have called "robust stability" throughout. Naturally, the larger ϵ, the more advantageous these two properties become. Unfortunately, it may not be possible to achieve $\tilde{\alpha}_\epsilon \leq 1$ if ϵ is chosen too big. Ideally thus, ϵ should be chosen as large as possible within the constraint that robust stability ($\tilde{\alpha}_\epsilon \leq 1$) can still be achieved. This tradeoff leads to a parameter-free version of the robust stabilization problem based on the smoothed spectral abscissa (Vanbiervliet et al., 2009), which is to find

$$\max_{(\epsilon,z)} \ \epsilon, \quad \text{subject to} \quad \tilde{\alpha}_\epsilon(\boldsymbol{W}(z)) \leq 1 \quad \text{and other constraints on } z \tag{4.19}$$

Note that the connectivity matrix is now called $\boldsymbol{W}(z)$ to make the dependence on the parameter set z more explicit. The constraints on z in problem 4.19 may for example express the fact that the inhibitory synapses cannot be positive, as in chapter 3. Let $(\epsilon^\star, z^\star)$ denote the solution to problem 4.19. Importantly, because $\tilde{\alpha}_\epsilon$ is a growing function of ϵ, the smoothed spectral abscissa in the optimum $\epsilon^\star$ must be exactly one, and $z^\star$ solves the minimization of $z \mapsto \tilde{\alpha}_{\epsilon^\star}(\boldsymbol{W}(z))$ subject to the constraints. Thus, the solution to problem 4.19 always involves

$$\mathcal{E}_0(\boldsymbol{W}(z^\star)) = \frac{1}{N\epsilon^\star} \tag{4.20}$$

Therefore, problem 4.19 which is about maximizing ϵ, is equivalent to the minimization of the expected evoked energy $\mathcal{E}_0(\boldsymbol{W}(z))$, subject to the constraints on z (or the minimization of $A(\boldsymbol{W}(z))$, by Equation 4.9). This observation was already formulated in Vanbiervliet et al. (2009), though with a different vocabulary.

In summary, robust stabilization can be done through the minimization of $\mathcal{E}_0(\boldsymbol{W})$, which by Equation 4.9 is equivalent to the minimization of the average variance $A(\boldsymbol{W})$ of the spontaneous firing rate fluctuations in single cells. I now show that, since $A(\boldsymbol{W})$ depends on quantities available "online" during spontaneous activity, a synaptic learning rule can be derived that approximates robust stabilization. This learning rule is structurally similar to that of Vogels et al. (2011).

4.3 A learning rule for approximate robust stabilization

I now consider a biologically feasible approximation to the robust stabilization problem (Equation 4.19), based on the idea of minimizing the intensity $A(\boldsymbol{W})$ of the spontaneous activity fluctuations under the linearized dynamics of Equation 4.2. The cost function to minimize

is

$$\mathcal{L}(\boldsymbol{W}) \quad = \quad \frac{1}{2}\sum_{i=1}^{N}\left\langle x_i^2(t)\right\rangle_t \tag{4.21}$$

where $\langle\cdot\rangle_t$ denotes temporal averaging. In general, since the network is recurrently connected, $\langle x_i^2(t)\rangle_t$ for postsynaptic neuron i depends upon all synaptic weights in the network, not only those that point to neuron i. This gives rise to a non-local learning rule. A straightforward local approximation would be to ignore such recurrent effects and assume that $\langle x_i^2(t)\rangle$ depends only on the i^{th} row of $\boldsymbol{W}$. In the case of linear dynamics (Equation 4.2 with $g(x) = x$), this yields the following approximate local gradient:

$$\frac{\partial \mathcal{L}_i}{\partial w_{ij}} \quad \simeq \quad \langle \hat{x}_j(t)\; x_i(t)\rangle_t \tag{4.22}$$

where $\hat{x}_j(t) \equiv \int_{-\infty}^{t} x_j(s)\exp[-(t-s)/\tau]ds$ reflects the low-pass filtering dynamics of unit i. Invoking standard stationarity and self-averaging arguments to drop the expectation brackets in Equation 4.22, one obtains a local and online inhibitory learning rule:

$$\frac{dw_{ij}}{dt} \quad \propto \quad -\hat{x}_j(t)\; x_i(t) \tag{4.23}$$

For inhibitory weights ($w_{ij} < 0$), this learning rule is of the Hebbian type, in the sense that correlated pre- (x_j) and post-synaptic (x_i) activities cause an increase in absolute synaptic efficacy.

I first test this local learning rule on the same inhibitory-stabilization task as in chapter 3. Since the theoretical considerations of section 4.2 apply only to the linear regime, I start from a random network of size $N = 200$ with an initial spectral abscissa of 0.9 so that the network is initially stable (Figure 4.1) (I extend to the nonlinear, initially unstable case later below). All excitatory weights remain fixed throughout, and the inhibitory weights are progressively modified as dictated by the learning rule, subject to i) a negativity constraint and ii) a sparsity constraint as in chapter 3 (only 40% of the I weights can be non-zero at a time). In short, I follow exactly the same method as described in section 3.4, except that steps 1–3 on page 75 are now replaced by the local and online synaptic update rule in Equation 4.23.

The primary effect of learning according to Equation 4.23 is indeed to decrease the spectral abscissa, which eventually saturates at a value of ~ 0.3. In comparison, the optimal learning rule (exact same procedure as in chapter 3 - ran on the exact same initial connectivity matrix – can reach a value of ~ 0.12 (Figure 4.1A and B). This discrepancy can be attributed to the crudeness of the locality approximation made to arrive at Equation 4.23, in a network that is fully recurrent. In any case, both learning procedures give rise to similar weight distributions (Figure 4.1C), and the individual synaptic weights that result from the two procedures are

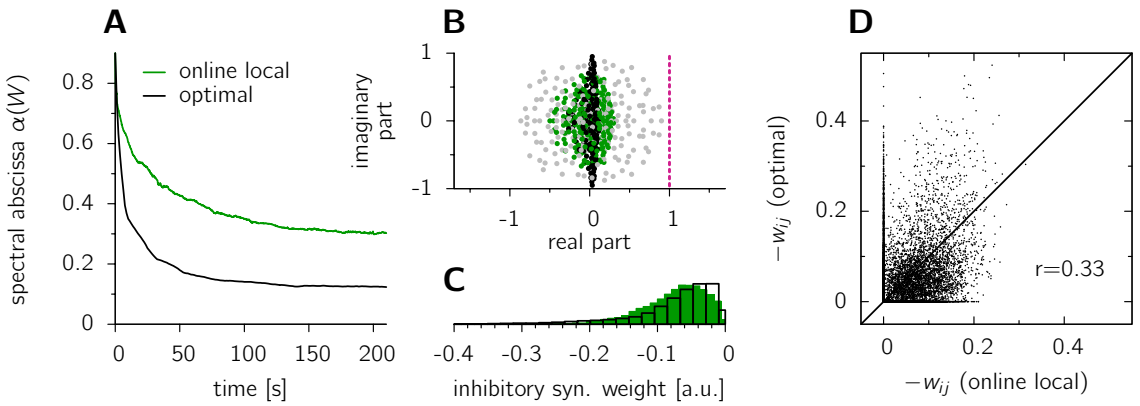


Figure 4.1: **Online learning rule for approximate optimal inhibitory stabilization**. (**A**) A stable random network with an initial spectral abscissa of 0.9 is subjected to the local inhibitory plasticity learning rule of Equation 4.23, aimed at minimizing the strength of activity fluctuations during spontaneous activity (modeled after Equation 4.2 with $g(x) = x$). Shown here is the evolution of the spectral abscissa as learning progresses (green). For the sake of comparison, we report the evolution of the spectral abscissa under optimal stability optimization (black; the x-axis should be interpreted as number of iterations of the optimization procedure; 200s $\simeq$ 160 iterations). (**B**) Eigenvalue spectrum of the connectivity matrix before (gray dots) and after learning (green dots). The black dots report the result of the optimal procedure. (**C**) Distribution of non-zero inhibitory synaptic efficacies after learning, with the same color scheme as in (B). Note that only 40% of the weights are non-zero. (**D**) The two procedures (online vs. optimal) yield correlated individual synaptic efficacies.

substantially correlated (Figure 4.1D), indicating that similar solutions are found by the two update rules.

Can we extrapolate to situations in which the network is initially unstable, yielding chaotic nonlinear dynamics? Let me again derive a local inhibitory plasticity learning rule, aimed at minimizing the amplitude of the spontaneous activity fluctuations. The network dynamics now obey the nonlinear version of Equation 4.2, with the nonlinearity $x \mapsto g(x)$ given by Equation 3.4 in chapter 3 (it is also sketched in Figure 3.3B). This function has maximum slope at $x = 0$, in which case the output is $\theta = 0$Hz (interpreted as the mean firing rate of each neuron). Since the variance of the spontaneous fluctuations may be "trivially" minimized by sending every neuron to a region where g saturates (corresponding to firing rates of either 0Hz or 100Hz), an homeostasis term must be added to the objective function to encourage firing rates to fluctuate around θ instead. The full cost function for postsynaptic neuron i now reads,

$$\mathcal{L}_i(\boldsymbol{W}) \quad = \quad \langle (y_i(t) - \langle y_i(t) \rangle_t)^2 \rangle_t + (\langle y_i(t) \rangle_t - \theta)^2 \tag{4.24}$$

with the notation $y_i(t) = g[x_i(t)]$. This reduces to

$$\mathcal{L}_i(\boldsymbol{W}) \quad = \quad \langle (y_i - \theta)^2 \rangle_t \tag{4.25}$$

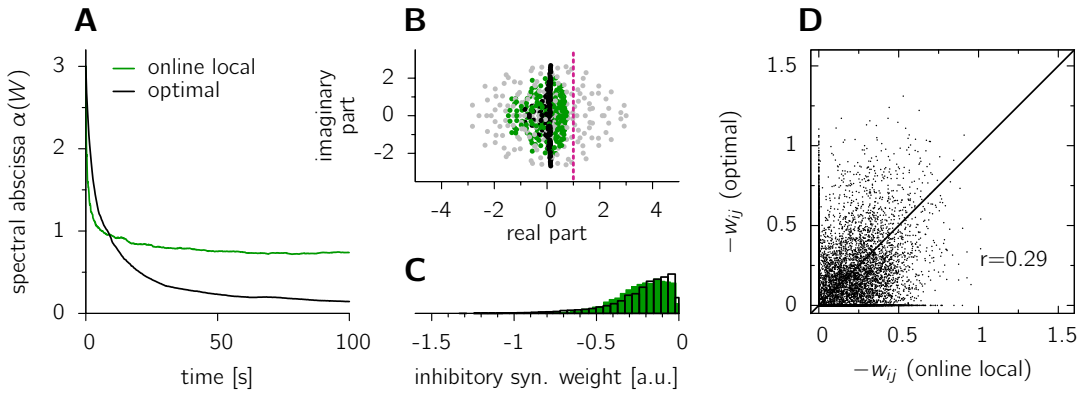


Figure 4.2: **Online learning in a nonlinear, initially unstable network**. Panels have the same meaning as in Figure 4.1. The initial connectivity matrix was a random balanced network with a spectral abscissa of 3. Learning ocurred following Equation 4.26 during spontaneous nonlinear dynamics (Equation 4.2).

Let us again make a local approximation, and assume that $x_i(t)$ depends only on the i^{th} row of $\boldsymbol{W}$. This yields the following gradient descent update rule:

$$\frac{dw_{ij}}{dt} \quad \propto \quad -g'[x_i(t)] \cdot \hat{y}_j(t) \cdot (y_i(t) - \theta) \tag{4.26}$$

where $\hat{y}_j(t) = \int_{-\infty}^{t} y_j(s) exp[-(t-s)/\tau] ds$. This learning rule successfully achieves linear stability in an initially unstable random balanced network with a spectral abscissa of 3 (Figure 4.2). Again the solution it finds is substantially correlated with the solution found by the optimal procedure based on the smoothed spectral abscissa. As expected though, the performance is substantially worse than that of the optimal stabilization strategy. When starting from a spectral abscissa of 10 as we did in chapter 3 (Figure 3.2), the online learning rule of Equation 4.26 is not able to achieve linear stability (not shown).

The learning rule of Equation 4.26 is structurally similar to the one Vogels and colleagues used in order to explain the maintenance of a detailed E/I balance in feedforward auditory pathways (Vogels et al., 2011). They also showed in recurrent network simulations that Hebbian learning at inhibitory synapses during spontaneous activity can suppress instabilities that may arise from strong and patterned excitatory feedback (destabilizing attractors). Here, through the link I have drawn with the smoothed spectral abscissa, I have provided a theoretical account for why a learning rule aimed at minimizing the strength of the spontaneous activity fluctuations can have such stabilizing effects.

In the simulations presented here, I have compared the inhibitory plasticity rule (Equation 4.26), supposed to approximate *robust* stabilization, to the control-theoretic stabilization procedure used in chapter 3. The latter was not strictly speaking an implementation of

the full robust stabilization problem in Equation 4.19; instead, it was a constrained minimization of $\tilde{\alpha}_\epsilon(\boldsymbol{W})$ with some heuristic choice of ϵ, and the main goal was to cause the spectral abscissa to decrease as much as possible. Future numerical studies should implement the full problem 4.19, and compare the results to the local learning rule in Equation 4.26. More work is also needed to check that the local plasticity learning rule derived here achieves *robust* stability in the sense that, after stabilization, the network remains stable despite comparatively large random perturbations of the synaptic weights.

CHAPTER 5

Stability in spatially structured networks via local inhibition

So far, this thesis has focussed on the issues of stability and amplification in local microcircuits, where the excitatory synaptic organization does not follow any obvious topology and has therefore been modelled as a random graph (chapter 2 and chapter 3). I have shown that stabilizing such microcircuits can be a difficult problem, and that inhibition must be finely tuned (chapter 3).

In this chapter, I zoom out and consider the cortical network at a larger scale. Specifically, I focus on the intrinsic lateral connections that form within the gray matter, and ramify up to several millimeters away from the soma of the presynaptic neuron. These originate from axon collaterals spreading toward distant columns (e.g. Gilbert and Wiesel (1983)). I look at the connectivity structure in a 3mm by 3mm patch of cortex (Figure 5.1A,B), and ignore its vertical (laminar) extent. A recent review has summarized the key qualitative features of synaptic organization at such a macroscopic scale into a canonical model (Voges et al., 2010) which I discuss here in terms of stability.

Perhaps surprisingly, I find that inhibitory feedback need not be complicated to stabilize recurrent excitation: it is sufficient i) that the average strength of the inhibitory connections be at least as large as that of the excitatory connections, and ii) that inhibitory synapses be confined to within a small radius around the soma of their presynaptic interneuron – smaller than the local radius of excitation. This situation seems to be in line with the anatomy of inhibition in the cortex (Voges et al., 2010; Packer and Yuste, 2011; Fino and Yuste, 2011).

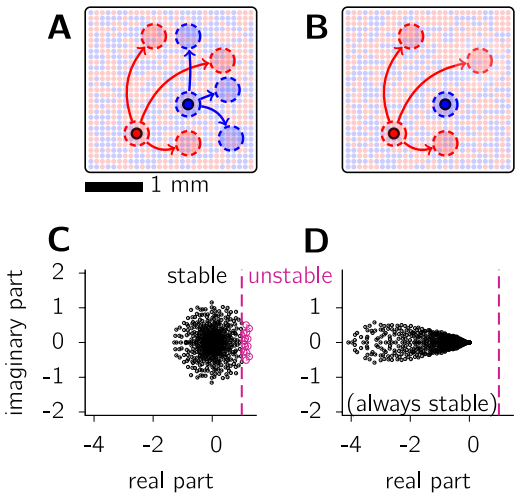


Figure 5.1: **Stability in spatially extended cortical networks.** (**A** and **B**) Schematics of network connectivities with topological organization within a 3mm by 3mm square patch of cortex. In both (A) and (B), excitatory (E, red) and inhibitory (I, blue) neurons make local connections onto neurons located within $\sim 300\mu m$ of their own cell bodies (dashed red and blue circles, around the corresponding highlighted E and I neurons). In (A), both E and I neurons also make long-range projections (arrows), clustered into 3 distant patches (dashed blue and red circles) placed at random positions for each presynaptic neuron (see text). In (B), inhibition stays purely local. In both cases, the average I connection is equal in magnitude to the average E connection (global balance). (**C** and **D**) Stability analysis of the corresponding connectivities in (A) and (B). The eigenvalues of each connectivity matrix are plotted in the complex plane. Purple points that lie to the right of the dashed purple vertical line represent unstable modes of network activity. The network with local dense inhibition is always stable, no matter how strong all the connections are overall.

5.1 Patchy model of macroscopic synaptic organization

In the class of network architectures considered here, all neurons – excitatory and inhibitory alike – establish a significant fraction of their outgoing synapses onto "neighbors", i.e. neurons located less than 300μm away from their soma (Figure 5.1A,B). This predominance of local connections reflects the physical proximity of neurons that are positioned less than 300μm apart, taking into account the spatial reach of their morphologies (Chklovskii, 2004; Douglas and Martin, 2007). Additionally, neurons may also project long-range, clustered connections to distant locations (Figure 5.1A,B, arrows). I assume 3 projection patches per neuron (the average given in Voges et al. (2010)), although variations in this number have little consequences on the results presented below. I consider two variants of the above synaptic architecture: *MODEL A* in which excitatory and inhibitory neurons all make local *and* long-range clustered projections (Figure 5.1A), and *MODEL B* in which inhibition remains local (Figure 5.1B).

The connectivities in *MODEL A* and *MODEL B* are idealized as dense connectivity matrices $\boldsymbol{W}$, where w_{ij} represents the probability that neuron j would synapse onto neuron i. The formalism to describe how $\boldsymbol{W}$ is computed is common to both models, only parameters differ. I assume a 25×25 grid on which both the excitatory neurons and inhibitory neurons are regularly positioned. Thus, there are $M = 625$ excitatory and $M = 625$ inhibitory cells, for a total of $N = 1250$ neurons in the network. Let $\boldsymbol{c}_i$ denote the position of neuron i on the

grid (vector of two relative coordinates between 0 and 1). For each neuron j, I pick a set of 3 target locations $\{\boldsymbol{\ell}_j(k), k = 1, 2, 3\}$, drawn randomly and uniformly on the grid. Like $\boldsymbol{c}_j$, $\boldsymbol{\ell}_j(k)$ is a pair of normalized spatial coordinates between 0 and 1. A connectivity matrix $\boldsymbol{W} = \{w_{ij}\}$ is then computed as

$$w_{ij} \cdot s_j = p_j^{\text{local}} \; F_{\sigma_j}\left[\Delta\left(\boldsymbol{c}_i|\boldsymbol{c}_j\right)\right] + (1 - p_j^{\text{local}}) \; \frac{1}{3}\sum_{k=1}^{3} F_{\sigma_j}\left[\Delta\left(\boldsymbol{c}_i|\boldsymbol{\ell}_j(k)\right)\right] \tag{5.1}$$

where $s_j = \pm 1$ determines the sign of the connection: positive for $j \leq M$ and negative otherwise. In Equation 5.1, $\Delta(\boldsymbol{a}|\boldsymbol{b})$ denotes the distance between position $\boldsymbol{a}$ and position $\boldsymbol{b}$ on the grid, assuming cyclic boundaries. The radial profile $F_{\sigma_j}(\Delta)$ – parameterized by a spread σ_j – expresses the decay of connection probability with distance, for connections that neuron j makes either around itself (first term in the r.h.s.) or around each of its 3 target locations (second term). Unless otherwise stated, I assume a Gaussian connectivity profile $F_{\sigma_j}(\Delta) \propto \exp\left(-\Delta^2/2\sigma_j^2\right)$. Note that function F_{σ_j} is further normalized so that $\int F_{\sigma_j} = 1$. Unless indicated otherwise, the spread σ_j is set to 300μm, independent of presynaptic neuron j.

Parameter p_j^{local} denotes the fraction of connections that neuron j makes in its local neighborhood, relative to its total number of outgoing synapses. For simplicity, I assume it depends only on unit j being excitatory or inhibitory: I thus define $p_E^{\text{local}} \equiv p_{j\leq M}^{\text{local}}$ and $p_I^{\text{local}} \equiv p_{j>M}^{\text{local}}$. *MODEL A* is thus characterized by $p_I^{\text{local}} = 0.5$ (equal proportion of local and long-range inhibition), while *MODEL B* assumes $p_I^{\text{local}} = 1$ (purely local inhibition). In both models I set $p_E^{\text{local}} = 0.5$ (Voges et al., 2010).

Finally, I assume that excitation and inhibition are globally balanced: the sum of all excitatory weights in matrix $\boldsymbol{W}$ is equal in absolute value to that of all inhibitory weights.

Note that the connectivity matrices are only defined up to a multiplicative constant here. For sufficiently weak connections, stability is not an issue. Here I wish to derive stability conditions that do not depend upon the overall connection strength. The stability arguments made below will be independent of the overall scaling of $\boldsymbol{W}$.

5.2 Stability via local inhibition

The key result here is the following: if inhibition is kept local (*MODEL B*), then a mere global E/I balance guarantees network stability (subject to some mild conditions as detailed below). As in previous chapters, stability is understood in the linear sense here: stability requires all the eigenvalues λ of $\boldsymbol{W}$ to lie within the left half-plane defined by $\text{Re}(\lambda) < 1$. Within the connectivity model of section 5.1 above, when inhibition is kept local and the

global E/I balance is satisfied, this stability condition is met, and robustly so: every real part in the spectrum of $\boldsymbol{W}$ is even negative (Figure 5.1D). This implies in particular that the overall absolute magnitude of the connection strengths can be made arbitrarily large without causing instability, so long as a global E/I balance holds.

There are a few additional conditions for network stability in *MODEL B*. First, the local radius of inhibition must be smaller than that of excitation (not shown). Second, the radial profile $F_{\sigma_j}(\Delta)$ of the local and distant patches around their respective centers (c.f. Equation 5.1) must satisfy a certain equation that we derive below, and which holds for the various types of distance-dependence reported in experiments (Hellwig, 2000; Perin et al., 2011; Levy and Reyes, 2012).

Importantly, the global E/I balance by itself is not a sufficient condition for stability. This is well illustrated by the qualitatively different behavior of *MODEL A*: despite a globally balanced mix of excitation and inhibition, the connectivity matrix has eigenvalues with positive real parts (Figure 5.1C), meaning that for sufficiently strong (though balanced) synaptic strengths instabilities are bound to develop. Thus, the overall strength of the synaptic efficacies in such a network must be carefully controlled in order to keep the dynamics stable, whereas model A does not require such fine tuning.

I now describe an attempt to characterize the stability of *MODEL B* analytically, and in particular to derive stability conditions on the spatial profile $F(\Delta)$. The connectivity matrix in *MODEL B* has the form

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_E & -\boldsymbol{W}_I \\ \boldsymbol{W}_E & -\boldsymbol{W}_I \end{pmatrix}, \tag{5.2}$$

Equation 5.2 expresses the assumption that, when a neuron targets a specific area, it makes connections to both E and I neurons in that local area with similar probabilities. Because $\boldsymbol{W}$ is made of two identical row blocks, it has at least $N/2$ zero eigenvalues (its rank is at most $N/2$). It is easily checked that the remaining $N/2$ eigenvalues of $\boldsymbol{W}$ are also the eigenvalues of $\boldsymbol{W}_0 = \boldsymbol{W}_E - \boldsymbol{W}_I$ with the same multiplicities (and see Supplemental Data of Murphy and Miller (2009)). Thus I will focus on $\boldsymbol{W}_0$.

Under what conditions does keeping inhibition local ensure that all eigenvalues of $\boldsymbol{W}_0$ (hence, of $\boldsymbol{W}$) have negative real parts? Characterizing the full spectrum of $\boldsymbol{W}$ is a difficult problem, but valuable insights can be gained from considering the reduced problem sketched in Figure 5.2A. In this toy problem, the 2D grid with cyclic boundaries is replaced by a continuous, infinite one-dimensional space. Thus, connectivity matrices become connectivity "kernels", and matrix multiplication becomes convolution with those kernels (see e.g. Dayan and Abbott (2001), chapter 7). Each excitatory neuron makes a single long-range projection patch at a distance p from itself. This distance is taken to be the same for every neuron, making

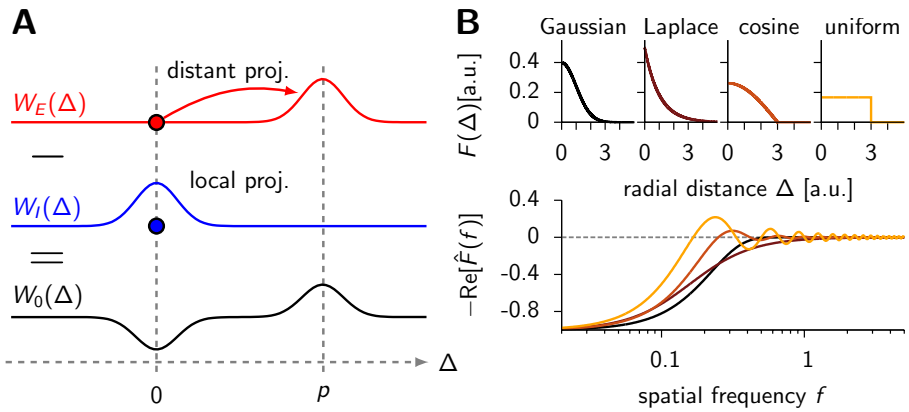


Figure 5.2: **Simplified version of MODEL B (local inhibition)**. (**A**) Schematics of the connectivity. Excitatory neurons (red circle) make a single projection patch at a distance of p from themselves (translation-invariant). Inhibitory neurons (blue circle) make a single local patch. Both types of projections (distant or local) have the same spatial profile $F(\Delta)$. Here it is sketched as a Gaussian profile, but I leave the shape free (see (B)) and seek conditions on $F(\Delta)$ so that every eigenvalue of the combined connectivity operator $W_0 = W_E - W_I$ (black) has negative real-part. (**B**) The eigenvalue real parts of the combined connectivity operator W_0 in (A) are bound to lie between 0 and $-\mathrm{Re}[\hat{F}(f)]$. These boundaries (bottom row) are plotted here for four different radial profiles $F(\Delta)$ (top row). For Gaussian-shaped or Laplacian distance-dependence, all eigenvalues have negative real parts, which corresponds to a robustly stable network.

---

the E connectivity translation-invariant. Each inhibitory neuron makes a single local patch (this is clearly also translation invariant). The spatial profile of each projection patch (local or distant) around its center is a positive and even function $F(\Delta)$ where Δ is the distance from center (Figure 5.2A). One may thus write the E connectivity from any point s in space to point $s + \Delta$ as as a kernel $W_E(\Delta) = F(\Delta - p)$, and similarly for the I connectivity with a kernel $W_I(\Delta) = F(\Delta)$. Because F is taken to be the same for both E and I projections, I need not model the local excitatory patch (present in the original Equation 5.1) explicitly, since it cancels out with half of the local inhibitory patch in the difference $W_0 = W_E - W_I$. Note that the connectivity in this toy problem satisfies the global E/I balance, as $\int W_0 = 0$.

An eigenvalue/eigenfunction pair $(\lambda, x(\cdot))$ of the convolution operator with the combined kernel $W_0(\Delta) = W_E(\Delta) - W_I(\Delta) = F(\Delta - p) - F(\Delta)$ satisfies

$$\int_{-\infty}^{+\infty} x(s - \Delta) W_0(\Delta) d\Delta = \lambda x(s) \qquad (\forall s \in \mathbb{R}) \tag{5.3}$$

Due to the assumed translation invariance, it is easy to show that Equation 5.3 is fulfilled by the Fourier modes $x_f(s) = e^{2j\pi f s}$ (with $j^2 = -1$). The eigenvalue associated with some spatial frequency f is $\lambda_f = \int W_0(\Delta) e^{-2j\pi f \Delta} d\Delta$, and the corresponding real part is given by

$$\mathrm{Re}(\lambda_f) = \int W_0(\Delta) \cos(2\pi f \Delta) d\Delta = \int [F(\Delta - p) - F(\Delta)] \cos(2\pi f \Delta) d\Delta \tag{5.4}$$

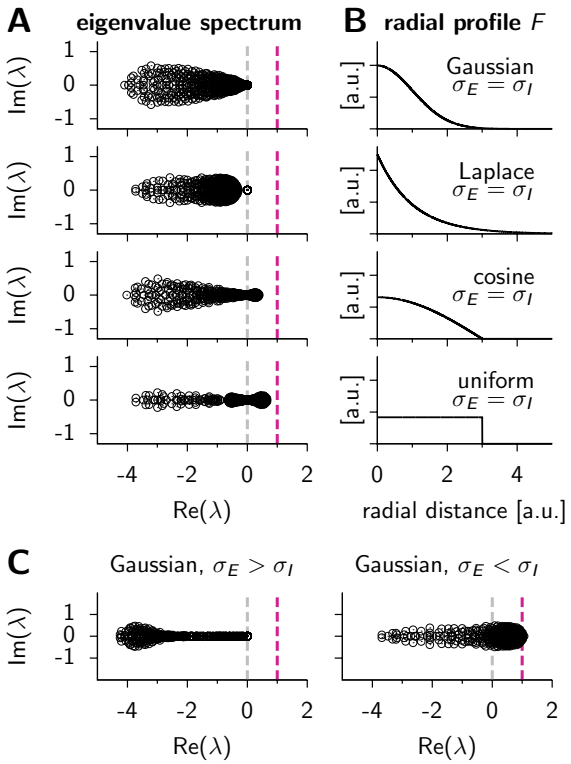


Figure 5.3: **Conditions for stability in networks with local inhibition and a global E/I balance** – The eigenvalue spectra of connectivity matrices are shown in (A), for the four types of connectivity distance-dependence shown in (B). Just as predicted by the simplified analysis of Figure 5.2, Gaussian and Laplacian radial profiles lead to robustly stable networks (all real parts in the spectrum are negative), while cosine-shaped or uniform profiles may cause instabilities if connections are strong enough. In (A), it is assumed that the spreads σ_E and σ_I of the E and I radial profiles are identical. In (C), we consider the remaining two cases (excitation broader, inhibition broader). For robust stability, excitation must be broader.

I now look at the extrema of $\mathrm{Re}(\lambda_f)$ as a function of p. Taking an arbitrary frequency f, and setting the derivative of $\mathrm{Re}(\lambda_f)$ w.r.t p to zero yields the extremum condition

$$\int F'(\Delta - p)\cos(2\pi f \Delta)d\Delta = 0 \tag{5.5}$$

In other words, the real part of the Fourier transform of $F'(\Delta + p)$ is zero, which can also be written as

$$\mathrm{Re}\left(2\pi j f \; e^{2\pi j p f}\hat{F}(f)\right) = 0 \tag{5.6}$$

where $\hat{F}$ is the Fourier transform of F. Since F is even, its Fourier transform is real. If $\hat{F}(f) \neq 0$ (which is the non-trivial case), the extremum condition in Equation 5.6 can therefore be written as $\sin(2\pi p f) = 0$, meaning $p = k/f$ or $p = (k + 1/2)/f$ for some integer k. The real part $\mathrm{Re}(\lambda_f)$ in Equation 5.4 is then easily evaluated at those extrema, and yields $\mathrm{Re}(\lambda_f) = 0$ in the first case, and $\mathrm{Re}(\lambda_f) = -\int F(\Delta)\cos(2\pi f \Delta)d\Delta = -\mathrm{Re}[\hat{F}(f)]$ in the second case. I conclude that $\mathrm{Re}(\lambda_f)$, seen as a function of p, lies somewhere between 0 and $-\mathrm{Re}[\hat{F}(f)]$, and is therefore negative if, and only if, $\mathrm{Re}[\hat{F}(f)] > 0$. Our robust stability condition may thus be expressed as

$$\mathrm{Re}[\hat{F}(f)] > 0 \qquad (\forall f > 0) \tag{5.7}$$

The Fourier transforms of both a Gaussian profile $F(\Delta) \propto \exp(-\Delta^2/2\sigma^2)$ and a Laplacian $F(\Delta) \propto \exp(-|\Delta|/\sigma)$ have positive real parts at all frequencies (Figure 5.2B). For those

radial profiles one therefore expects a balanced network with local inhibition to be stable irrespective of the absolute strength of the connections. This however is not true of other radial profiles, such as a rectified cosine $F_{\sigma_j}(\Delta) \propto \left[\cos\left(\Delta\pi/6\sigma_j\right)\right]_+$ or a uniform flat profile on the interval $[-3\sigma_j, 3\sigma_j]$ (Figure 5.2B).

Despite numerous simplifying assumptions, the condition in Equation 5.7 is predictive of stability even in the full model of section 5.1. Numerically, Equation 5.7 appears to be an accurate sufficient condition for stability via local inhibition, even when the excitatory neurons make both local and multiple long-range projections, that are not translation-invariant, and formed in 2D space. This is shown in Figure 5.3A, where I plot the eigenvalue spectra of the matrices built as described in section 5.1 with different forms of distance-dependence $F(\Delta)$ shown in Figure 5.3B.

Finally, a derivation similar to the above can be made to show that the local radius of inhibition must be smaller than that of excitation for all the eigenvalues of $\boldsymbol{W}$ to have negative real parts (Figure 5.3C). This is opposite to the well-known "mexican-hat" connectivity structure, known to yield bump attractor dynamics when combined with a saturating input-output gain function in single neurons (Ben-Yishai et al., 1995).

CHAPTER 6

STDP in adaptive neurons gives close-to-optimal information transmission

This chapter is largely unrelated to the first three chapters of this thesis, and is therefore difficult to introduce in their context. I let it stand alone as a contribution to the field of synaptic plasticity, and let the abstract and introduction of the paper position the work within the previous literature on spike timing-dependent plasticity (STDP). This work was published in

> STDP in adaptive neurons gives close-to-optimal information transmission
> **Guillaume Hennequin**, Wulfram Gerstner and Jean-Pascal Pfister
> *Frontiers in Computational Neuroscience* (2010)

Abstract

Spike-frequency adaptation is known to enhance the transmission of information in sensory spiking neurons by rescaling the dynamic range for input processing, matching it to the temporal statistics of the sensory stimulus. Achieving maximal information transmission has also been recently postulated as a role for Spike-Timing Dependent Plasticity (STDP). However, the link between optimal plasticity and STDP in cortex remains loose, as does the relationship between STDP and adaptation processes. We investigate how STDP, as described by recent minimal models derived from experimental data, influences the quality of information transmission in an adapting neuron. We show that a phenomenological model based on triplets of spikes yields almost the same information rate as an optimal model specially designed to this end. In contrast, the standard pair-based model of STDP does not improve information transmission as much. This result holds not only for additive STDP with hard weight bounds, known to produce bimodal distributions of synaptic weights, but also for weight-dependent STDP in the context of unimodal but skewed weight distributions. We analyze the similarities between the triplet model and the optimal learning rule, and find that the triplet effect is an important feature of the optimal model when the neuron is adaptive. If STDP is optimized for information transmission, it must take into account the dynamical properties of the postsynaptic cell, which might explain the target-cell specificity of STDP. In particular, it accounts for the differences found *in vitro* between STDP at excitatory synapses onto principal cells and those onto fast-spiking interneurons.

6.1 Introduction

The experimental discovery of Spike Timing-Dependent Plasticity (STDP) in the mid-nineties (Markram et al., 1997; Bell et al., 1997; Magee and Johnston, 1997; Bi and Poo, 1998; Zhang et al., 1998) led to two questions, in particular. The first is: what is the simplest way of describing this complex phenomenon? This question has been answered in a couple of minimal models (phenomenological approach) whereby long-term potentiation (LTP) and long-term depression (LTD) are reduced to the behavior of a small number of variables (Gerstner et al., 1996; Kempter et al., 1999; Song et al., 2000; van Rossum et al., 2000; Rubin et al., 2001; Gerstner and Kistler, 2002a; Pfister and Gerstner, 2006; Froemke et al., 2006; Clopath et al., 2010) – see Morrison et al. (2008) for a review. Because they are inspired by *in vitro* plasticity experiments, the state variables usually depend solely on what is experimentally controlled, i.e. on spike times and possibly on the postsynaptic membrane potential. They are computationally cheap enough to be used in large-scale simulations (Morrison et al., 2007; Izhikevich and Edelman, 2008). The second question has to do with the functional

role of STDP: what is STDP good for? The minimal models mentioned above can address this question only indirectly, by solving the dynamical equation of synaptic plasticity for input with given stationary properties (Kempter et al., 1999; van Rossum et al., 2000; Rubin et al., 2001). An alternative approach is to postulate a role for synaptic plasticity, and formulate it in the mathematical framework of optimization ("top-down approach"). Thus, in artificial neural networks, Hebbian-like learning rules were shown to arise from unsupervised learning paradigms such as principal components analysis (Oja, 1982, 1989), independent components analysis (Intrator and Cooper, 1992; Bell and Sejnowski, 1995; Clopath et al., 2008), maximization of mutual information (Linsker, 1989), sparse coding (Olshausen and Field, 1996; Smith and Lewicki, 2006) and predictive coding (Rao and Ballard, 1999). In spiking neurons, local STDP-like learning rules were obtained from optimization criteria such as maximization of information transmission (Chechik, 2003; Toyoizumi et al., 2005, 2007), information bottleneck (Klampfl et al., 2009), maximization of the neuron's sensitivity to the input (Bell and Parra, 2005), reduction of the conditional entropy (Bohte and Mozer, 2007), slow-feature analysis (Sprekeler et al., 2007), and maximization of the expected reward (Xie and Seung, 2004; Pfister et al., 2006; Florian, 2007; Sprekeler et al., 2009).

The functional consequences of STDP have mainly been investigated in simple integrate-and-fire neurons, where the range of temporal dependencies in the postsynaptic spike train spans no more than the membrane time constant. Few studies have addressed the question of the synergy between STDP and more complex dynamical properties on different timescales. In Seung (2003), more complex dynamics were introduced not at the cellular level, but through short-term plasticity of the synapses. The postsynaptic neuron was then able to become selective to temporal order in the input. Another elegant approach to this question was taken in Lengyel et al. (2005) in a model of hippocampal autoassociative memory. Memories were encoded in the phase of firing of a population of neurons relative to an ongoing theta oscillation. Under the assumption that memories are stored using a classical form of STDP, they derived the form of the postsynaptic dynamics that would optimally achieve their recall. This turned out to match what they recorded *in vitro*, suggesting that STDP might optimally interact with the dynamical properties of the postsynaptic cell in this memory storage task.

More generally, optimality models are ideally suited to study plasticity and dynamics together. Indeed, optimal learning rules contain an explicit reference to the dynamical properties of the postsynaptic cell, by means of the transfer function that maps input to output values. This function usually appears in the formulation of a gradient ascent on the objective function. In this article, we exploit this in order to relate STDP to Spike-Frequency Adaptation (SFA), an important feature of the dynamics of a number of cell types found in cortex. Recent phenomenological models of STDP have emphasized the importance of the interaction between postsynaptic spikes in the LTP process (Senn et al., 2001; Pfister and Gerstner, 2006;

Clopath et al., 2010). In these models, the amount of LTP obtained from a pre-before-post spike pair increases with the number of postsynaptic spikes fired in the recent past, which we call the "triplet effect" (combination of 1 pre-spike and at least 2 post-spikes). The timescale of this post-post interaction was fitted to *in vitro* STDP experiments, and found to be very close to that of adaptation (100 to 150 ms).

We reason that STDP may be ideally tuned to SFA of the postsynaptic cell. We specifically study this idea within the framework of optimal information transmission (infomax) between input and output spike trains. We compare the performance of a learning rule derived from the infomax principle in Toyoizumi et al. (2005), to that of the triplet model developed in Pfister and Gerstner (2006). We also compare them to the standard pair-based learning window used in most STDP papers. Performance is measured in terms of information theoretic quantities. We find that the triplet learning rule yields a better performance than pair-STDP on a spatio-temporal receptive field formation task, and that this advantage crucially depends on the presence of postsynaptic SFA. This reflects a synergy between the triplet effect and adaptation. The reasons for this optimality are further studied by showing that the optimal model features a similar triplet effect when the postsynaptic neuron adapts. We also show that both the optimal and triplet learning rules increase the variability of the postsynaptic spike trains, and enlarge the frequency band in which signals are transmitted, extending it towards lower frequencies (1-5 Hz). Finally, we exploit the optimal model to predict the form of the STDP mechanism for two different target cell types. The results agree qualitatively with the *in vitro* data reported for excitatory synapses onto principal cells and those onto fast-spiking inhibitory interneurons. In the model, the learning windows are different because the intrinsic dynamical properties of the two postsynaptic cell types are different. This might be the functional reason for the target-cell specificity of STDP.

6.2 Material and Methods

6.2.1 Neuron model

We simulate a single stochastic point neuron (Gerstner and Kistler, 2002b) and a small portion of its incoming synapses ($N = 1$ for the simulation of *in vitro* experiments, $N = 100$ in the rest of the paper). Each postsynaptic potential (PSP) adds up linearly to form the total modeled synaptic drive

$$u(t) = \sum_{j=1}^{N} w_j \epsilon_j(t) \tag{6.1}$$

with

$$\epsilon_j(t) = \int_0^t x_j(t') \exp\left(-\frac{t-t'}{\tau_m}\right) dt' \tag{6.2}$$

where $x_j(t) = \sum_{t_j^f} \delta\left(t - t_j^f\right)$ denotes the j^{th} input spike train, and w_j (mV) are the synaptic weights. The effect of thousands of other synapses is not modeled explicitly, but treated as background noise. The firing activity of the neuron is entirely described by an instantaneous firing density

$$\rho(t) = g[u(t)]M(t) \tag{6.3}$$

where

$$g[u] = g_0 + r_0 \log\left[1 + \exp\left(\beta\left(u - u_T\right)\right)\right] \tag{6.4}$$

is the gain function, drawn in Figure 6.1A. Refractoriness and SFA both modulate the instantaneous firing rate via

$$M(t) = \exp\left[-\left(g_R(t) + g_A(t)\right)\right] \tag{6.5}$$

The variables g_R and g_A evolve according to

$$\frac{dg_R}{dt} = -\frac{g_R(t)}{\tau_R} + q_R\, y(t) \qquad \text{and} \qquad \frac{dg_A}{dt} = -\frac{g_A(t)}{\tau_A} + q_A\, y(t) \tag{6.6}$$

where $y(t) = \sum_{t_{\text{post}}^f} \delta\left(t - t_{\text{post}}^f\right)$ is the postsynaptic spike train and $0 < \tau_R \ll \tau_A$ are the time constants of refractoriness and adaptation respectively. The firing rate thus becomes a compressive function of the average gain, as shown in Figure 6.1B. The response of the neuron to a step in input firing rate is depicted in Figure 6.1C.

For the simulation of *in vitro* STDP experiments, only one synapse is investigated. The potential u is thus given a baseline u_b (to which the PSP of the single synapse will add) such that $g(u_b)$ yields a spontaneous firing rate of 7.5 Hz (Figure 6.1B).

In some of our simulations, postsynaptic SFA is switched off ($q_A = 0$). In order to preserve the same average firing rate given the same synaptic weights, r_0 is rescaled accordingly (Figure 6.1A and Figure 6.1B, dashed lines).

In the simulation of Figure 6.8, we add a third variable g_B in the after-spike kernel M in order to model a fast-spiking inhibitory interneuron. This variable jumps down ($q_B < 0$) following every postsynaptic spike, and decays exponentially with time constant τ_B (with $\tau_R \ll \tau_B < \tau_A$).

All simulations were written in Objective Caml and run on a standard desktop computer operated by Linux. We used simple Euler integration of all differential equations, with 1 ms time resolution (0.1 ms for the simulation of *in vitro* experiments). All parameters are listed in Table 6.1 together with their values.

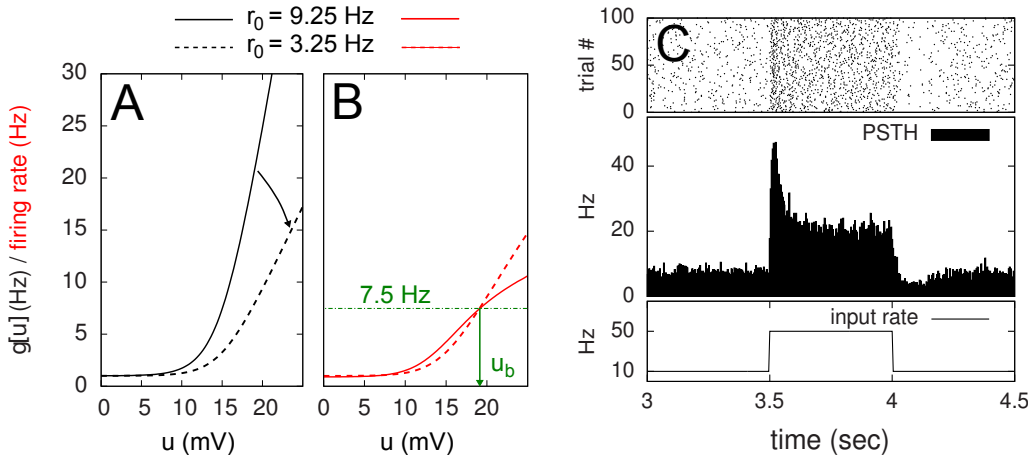


Figure 6.1: **Stochastic neuron model**. **(A)** The gain function $g(u)$ (Equation 6.4, solid line here) shows the momentary rate of a non-refractory neuron as a function of the membrane potential u. **(B)** The mean rate $\langle g[u(t)]M(t)\rangle$ of a neuron with refractoriness and adaptation is lower (solid red line). The baseline potential u_b used in the simulation is defined as the membrane potential that yields a spontaneous firing rate of 7.5 Hz (green arrow and dashed line). In some simulations, we need to switch off adaptation, but we want the same holding potential u_b to evoke the same 7.5 Hz output firing rate. The slope r_0 of the gain function is therefore rescaled (**(A)**, dashed curve) so that the frequency curves in the adaptation and no-adaptation cases (**(B)**, solid and dashed red curves) cross at $u = u_b$. **(C)** Example response property of an adaptive neuron. A single neuron receives synaptic inputs from 100 poisson spike trains with a time-varying rate. The experiment is repeated 1000 times independently. Bottom: the input rate jumps from 10 to 50 Hz, stays there for half a second and returns back to 10 Hz (bottom). Middle: Persi-Stimulus Time Histogram (PSTH, 4 ms bin). Top: example spike trains (first 100 trials).

6.2.2 Presynaptic firing statistics

To analyze the evolution of information transmission under different plasticity learning rules, we consider $N = 100$ periodic input spike of 5 seconds duration generated once and for all (see below). This "frozen noise" is then replayed continuously, feeding the postsynaptic neuron for as long as is necessary (e.g. for learning, or for mutual information estimation).

To generate the time-varying rates of the N processes underlying this frozen noise, we first draw point events at a constant Poisson rate of 10 Hz, and then smooth them with a Gaussian kernel of width 150 ms. Rates are further multiplicatively normalized so that each presynaptic neuron fires an average of 10 spikes per second. We emphasize that this process describes the statistics of the inputs *across different learning experiments*. When we mention "independent trials", we mean a set of experiments which have their own independent realizations of those input spike trains. However, in one learning experiment, a single such set of N input spike trains is chosen and replayed continuously as input to the postsynaptic neuron. The input is therefore deterministic and periodic. When the periodic input is generated, some neurons

can happen to fire at some point during those 5 seconds within a few milliseconds of each other, and by virtue of the periodicity, these synchronous firing events will repeat in each period, giving rise to strong spatio-temporal correlations in the inputs. We are interested in seeing how different learning rules can exploit this correlational structure to improve the information carried by the postsynaptic activity about those presynaptic spike trains. We now describe what we mean by information transmission under this specific stimulation scenario.

6.2.3 Information theoretic measurements

The neuron can be seen as a noisy communication channel in which multidimensional signals are compressed and distorted before being transmitted to subsequent receivers. The goodness of a communication channel is traditionally measured by Shannon's mutual information between the input and output variables, where the input is chosen randomly from some "alphabet" or vocabulary of symbols.

Here, the input is deterministic and periodic (Figure 6.2A). We therefore define the quality of information transmission by the reduction of uncertainty about the phase of the current input if we observe a certain output spike train at an unknown time. In discrete time (with time bin $\Delta = 1$ms), there are only $N_\phi = 5000$ possible phases since the input has a period of 5 seconds. Therefore, the maximum number of bits that the noisy postsynaptic neuron can transmit is $\log_2(N_\phi) \simeq 12.3$ bits. We further assume that an observer of the output neuron can only see "words" corresponding to spike trains of finite duration $T = K\Delta$. We assume $T = 1$ second for most of the paper, which corresponds to $K = 1000$ time bins. This choice is justified below.

The discretized output spike trains of size K (binary vectors), called $\mathbf{Y}^K$, can be observed at random times and plays the role of the output variable. The input random variable is the phase $\mathbf{\Phi}$ of the input. The quality of information transmission is quantified by the mutual information, i.e. the difference between the total response entropy $H(\mathbf{Y}^K) = \left\langle \log_2 P(\mathbf{Y}^K) \right\rangle_{\mathbf{Y}^K}$ and the noise entropy $H(\mathbf{Y}^K|\mathbf{\Phi}) = \left\langle \left\langle \log_2 P(\mathbf{Y}^K|\mathbf{\Phi}) \right\rangle_{\mathbf{Y}^K|\mathbf{\Phi}} \right\rangle_{\mathbf{\Phi}}$. Here $\langle\cdot\rangle$ denotes the ensemble average. In order to compute these entropies, we need to be able to estimate the probability of occurence of any sample word Y^K, knowing and not knowing the phase. To do so, a large amount of data is first generated. The noisy neuron is fed continously for a large number of periods $N_p = 100$ with a single periodic set of input spike trains and a fixed set of synaptic weights. The output spikes are recorded with $\Delta = 1$ms precision. From this very long output spike train, we randomly pick words of length K and gather them in a set $\mathcal{S}$. We take $|\mathcal{S}| = 1000$. This is our sample data.

In general, estimating the probability of a random binary vector of size K is very difficult if

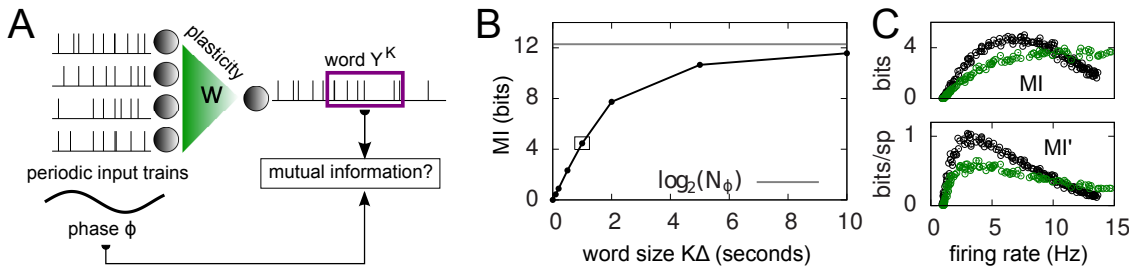


Figure 6.2: **Information transmission through a noisy postsynaptic neuron**. **(A)** Schematic representation of the feedforward network. 5-second input spike trains repeat continuously in time (periodic input) and drive a noisy and possibly adapting output neuron via plastic synapses. It is assumed that an observer of the output spike train has access to portions Y^K of it, called "words", of duration $T = K\Delta$. The observer does not have access to a clock, and therefore has a flat prior expectation over possible phases before observing a word. The goodness of the system, given a set of synaptic weights $\mathbf{w}$, is measured by the reduction of uncertainty about the phase, gained from the observation of an output word Y^K (mutual information, see text). **(B)** For a random set of synaptic weights (20 weights at 4mV, the rest at zero), the mutual information (MI) is reported as a function of the output word size $K\Delta$. Asymptotically, the MI converges to the theoretical limit given by $\log_2(N_\phi) \simeq 12.3$ bits. In the rest of this study, 1-second output words are considered (square). **(C)** Mutual information (MI, top) and information per spike (MI', bottom) as a function of the average firing rate. Black: with SFA. Green: without SFA. Each dot is obtained by setting a fraction of randomly chosen synaptic efficacies to the upper bound (4 mV) and the rest to 0. The higher the fraction of non-zero weights, the higher the firing rate. The information per spike is a relevant quantity because spike generation costs energy.

K is large. Luckily, we have a statistical model for how spike trains are generated (Equation 6.3), which considerably reduces the amount of data needed to produce a good estimate. Specifically, if the refractory state of the neuron $[g_R(t), g_A(t)]$ is known at time t (initial conditions), then the probability $1 - \exp(-\rho_k\Delta) \simeq \rho_k\Delta$ of the postsynaptic neuron spiking is also known for each of the K time bins following t (Equation 6.3 to Equation 6.5). The neuron model gives us the probability that a word Y^K occured at time t – not necessarily the time at which the word was actually picked – (Toyoizumi et al., 2005):

$$P\left(Y^K | \, t, g_R(t), g_A(t)\right) = \exp\left[\sum_{k=1}^{K} Y_k^K \log\left(\rho_k\Delta\right) + \left(1 - Y_k^K\right) \log\left(1 - \rho_k\Delta\right)\right] \tag{6.7}$$

where $\rho_k = \rho(t + k\Delta)$ and Y_k^K is one if there is a spike in the word at position k, and zero otherwise. To compute the conditional probability of occurence of a word Y^K knowing the phase ϕ, we have to further average Equation 6.7:

$$P(Y^K|\phi) = \left\langle P(Y^K|t)\right\rangle_{t \text{ with } \Phi(t)=\phi} \tag{6.8}$$

where $\Phi(t) = 1 + (t \bmod N_\phi)$ denotes the phase at time t. Averaging over multiple times with same phase also averages over the initial conditions $[g_R(t), g_A(t)]$, so that they do not

appear in Equation 6.8. The average in Equation 6.8 is estimated using a set of 10 randomly chosen times t_i with $\Phi(t_i) = \phi$.

The full probability of observing a word Y^K is given by $P(Y^K) = \frac{1}{N_\phi} \sum_{\phi=1}^{N_\phi} P(Y^K|\phi)$ where $P(Y^K|\phi)$ is computed as described above. Owing to the knowledge of the model that underlies spike generation, and to this huge averaging over all the possible phases, the obtained $P(Y^K)$ is a very good estimate of the true density. We can then take a Monte-Carlo approach to estimate the entropies, using the set $\mathcal{S}$ of randomly picked words: $H(\mathbf{Y}^K) = -\sum_{Y^K} P(Y^K) \log_2 P(Y^K)$ can be estimated by

$$\hat{H}(\mathbf{Y}^K) = -\frac{1}{|\mathcal{S}|} \sum_{Y^K \in \mathcal{S}} \log_2 P(Y^K) \tag{6.9}$$

and $H(\mathbf{Y}^K|\mathbf{\Phi}) = -\sum_\phi P(\phi) \sum_{Y^K} P(Y^K) \frac{P(Y^K|\phi)}{P(Y^K)} \log_2 P(Y^K|\phi)$ is estimated using

$$\hat{H}(\mathbf{Y}^K|\mathbf{\Phi}) = -\frac{1}{N_\phi} \sum_{\phi=1}^{N_\phi} \frac{1}{|\mathcal{S}|} \sum_{Y^K \in \mathcal{S}} \frac{P(Y^K|\phi)}{P(Y^K)} \log_2 P(Y^K|\phi) \tag{6.10}$$

The mutual information (MI) estimate is the difference of these two entropies, and is expressed in bits. In Figure 6.2C, we introduce the information per spike MI' (bits/spike), obtained by dividing the MI by the expected number of spikes in a window of duration $K\Delta$. Figure 6.2B shows that the MI approaches its upper bound $\log_2(N_\phi)$ as the word size increases. The word size considered here (1 second) is large enough to capture the effects of SFA while being small enough not to saturate the bound.

Although we constrain the postsynaptic firing rate to lie around a fixed value ρ_{targ} (see homeostasis in the next section), the rate will always jitter. Even a small jitter of less than 0.5 Hz (which we have in the present case) makes it impossible to directly compare entropies across learning rules. Indeed, while the mutual information depends only weakly on small deviations of the firing rate around ρ_{targ}, the response and noise entropies have much larger (co-)variations. In order to compare the entropies across learning rules, we need to know what the entropy would have been if the rate was exactly ρ_{targ} instead of $\rho_{\text{targ}} + \varepsilon$. We therefore compute the entropy ($H(\mathbf{Y}^K)$ or $H(\mathbf{Y}^K|\mathbf{\Phi})$) for different firing rates in the vicinity of ρ_{targ}. These firing rates are achieved by slightly rescaling the synaptic weights, i.e. $w_{ij} \leftarrow \kappa w_{ij}$ where κ takes several values around 1. We then fit a linear model $H = a\rho + b$, and evaluate H at ρ_{targ}.

The computation of the conditional probabilities $P(Y^K|\phi)$ was accelerated on an ATI Radeon (HD 4850) graphics processing unit (GPU), which was 130 times faster than a decent CPU implementation.

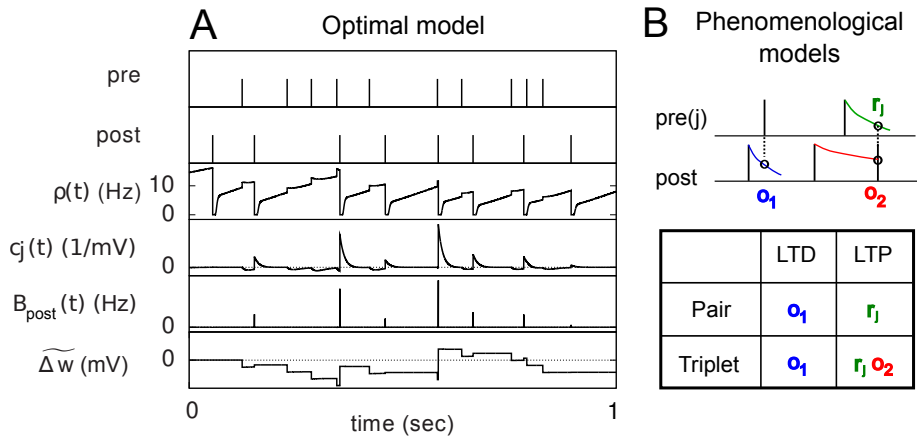


Figure 6.3: **Description of the three learning rules. (A)** Time course of the variables involved in the optimal model. $\widetilde{\Delta w}$ denotes the cumulative weight change. **(B)** Schematic representation of the phenomenological models of STDP used in this paper. Each presynaptic spike yields LTD proportionally to o_1 (blue trace) in both models (pair and triplet). In the pair model, postsynaptic spikes evoke LTP proportionally to r_j (green trace), while in the triplet model r_j is combined with an additional postsynaptic trace o_2 (red).

6.2.4 Learning rules

6.2.4.1 Optimal learning rule

The optimal learning rule aims at maximizing information transmission under some metabolic constraints ("infomax" principle). Toyoizumi et al. (2005) and Toyoizumi et al. (2007) showed that this can be achieved my means of a stochastic gradient ascent on the following objective function

$$\mathcal{L} = \mathcal{I} - \gamma\mathcal{D} - \lambda\Psi \tag{6.11}$$

whereby the mutual information $\mathcal{I}$ between input and output spike trains competes with a homeostatic constraint on the mean firing rate $\mathcal{D}$ and a metabolic penalty Ψ for strong weights that are often active. The first constraint is formulated as $\mathcal{D} = \mathrm{KL}\left[P(\mathbf{Y}^K), \tilde{P}(\mathbf{Y}^K)\right]$ where KL denotes the Kullback-Leibler (KL) divergence. P denotes the true probability distribution of output spike trains produced by the stochastic neuron model, while $\tilde{P}$ assumes a similar model in which the gain $g(t)$ is kept constant at a target gain g_{targ}. Minimizing the divergence between P and $\tilde{P}$ therefore means driving the average gain close to g_{targ}, thus implementing firing rate homeostasis. The second constraint reads $\Psi = \sum_j w_j \left\langle n_j \right\rangle_{X_K}$, whereby the cost for synapse j is proportional to its weight w_j and to the average number n_j of presynaptic spikes relayed during the K time bins under consideration. The Lagrange multipliers γ and λ set the relative importance of the three objectives.

Performing gradient ascent on $\mathcal{L}$ yields the following online learning rule (Toyoizumi et al.,

2005, 2007):

$$\frac{dw_j}{dt} = \eta_o \left[C_j(t) B_{\text{post}}(t) - \lambda x_j(t) \right] \tag{6.12}$$

where

$$C_j(t) = \int_0^t dt' \exp\left(-\frac{t-t'}{\tau_C}\right) \epsilon_j(t') \frac{g'[u(t')]}{g[u(t')]} \left[y(t') - g[u(t')] M(t') \right] \tag{6.13}$$

and

$$B_{\text{post}}(t) = y(t) \log\left[\frac{g[u(t)]}{\bar{g}} \left(\frac{g_{\text{targ}}}{\bar{g}} \right)^{\gamma} \right] - M(t) \left[g[u(t)] - \bar{g} + \gamma \left(g_{\text{targ}} - \bar{g} \right) \right] \tag{6.14}$$

η_o is a small learning rate. The first term C_j is Hebbian in the sense that it reflects the correlations between the input and output spike trains. B_{post} is purely postsynaptic: it compares the instantaneous gain g to its average $\bar{g}$ (information term), as well as the average gain to its target value g_{targ} (homeostasis). The average $\bar{g}$ is estimated online by a low pass filter of g with time constant τ_g. The time course of these quantities is shown in Figure 6.3A for example spike trains of 1 second duration, for $\gamma = 0$.

Because of the competition between the three objectives in Equation 6.11, the homeostatic constraint does not yield the exact desired gain g_{targ}. In practice, we set the value of g_{targ} empirically, such that the actual mean firing rate approaches the desired value.

Finally, we use τ_C, η_o and λ as three free parameters to fit the results of *in vitro* STDP pairing experiments (Figure 6.8). τ_C is set empirically equal to the membrane time constant $\tau_M = 20$ ms, while η_o and λ are determined through a least-squares fit of the experimental data. The learning rate η_o can be rescaled arbitrarily. In the simulations of receptive-field development (Figure 6.4,Figure 6.5, and Figure 6.6), λ is set to zero so as not to perturb unnecessarily the prime objective of maximizing information transmission. It is also possible to remove the homeostasic constraint ($\gamma = 0$) in the presence of SFA. As can be seen in Figure 6.2C, the MI has a maximum at 7.5Hz when the neuron adapts, so that firing rate control comes for free in the information maximization objective. We therefore set $\gamma = 0$ when the neuron adapts, and $\gamma = 1$ when is does not. In fact, the homeostasis constraint only slightly impairs the infomax objective: we have checked that the MI reached after learning (Figure 6.4 and Figure 6.5) does not vary by more than 0.1 bit when γ takes values as large as 20.

6.2.4.2 Triplet-based learning rule

We use the minimal model developed in Pfister and Gerstner (2006) with "all-to-all" spike interactions. Presynaptic spikes at synapse j leave a trace r_j (Figure 6.3B) which jumps by

1 after each spike and otherwise decays exponentially with time constant τ_+. Similarly, the postsynaptic spikes leave two traces, o_1 and o_2, which jump by 1 after each postsynaptic spike and decay exponentially with time constants τ_- and τ_y respectively:

$$\frac{dr_j}{dt} = -\frac{r_j}{\tau_+} + x_j(t) \qquad \frac{do_1}{dt} = -\frac{o_1}{\tau_-} + y(t) \qquad \frac{do_2}{dt} = -\frac{o_2}{\tau_y} + y(t) \tag{6.15}$$

where $x_j(t)$ and $y(t)$ are sums of δ-functions at each firing time as introduced above. The synaptic weight w_j undergoes LTD proportionally to o_1 after each presynaptic spike, and LTP proportionally to $r_j o_2$ following each postsynaptic spike:

$$\frac{dw_j}{dt} = \eta_3 \left[A_3^+ r_j(t) o_2(t-\varepsilon) y(t) - A_2^- o_1(t) x_j(t) \right] \tag{6.16}$$

where η_3 denotes the learning rate. Note that o_2 is taken just before its update. Under the assumption that pre- and postsynaptic spike trains are independent Poisson processes with rates ρ_x and ρ_y respectively, the average weight change was shown in Pfister and Gerstner (2006) to be proportional to

$$\langle \Delta w \rangle \propto \rho_x \rho_y \left(\rho_y - \frac{\tau_- A_2^-}{\tau_+ \tau_y A_3^+} \right) \tag{6.17}$$

The rule is thus structurally similar to a BCM learning rule (Bienenstock et al., 1982) since it is linear in the presynaptic firing rates and nonlinear in the postsynaptic rate. It is possible to roughly stabilize the postsynaptic firing rate at a target value ρ_{targ}, by having A_2^- slide in an activity-dependent manner:

$$A_2^-(t) = \tilde{A}_2^- \frac{\bar{\rho}^3(t)}{\rho_{\text{targ}}^3} \tag{6.18}$$

where $\tilde{A}_2^-$ is a starting value and $\bar{\rho}$ is an average of the instantaneous firing rate on the timescale of seconds or minutes (time constant τ_ρ). Finally, A_3^+ is set to make ρ_{targ} an initial fixed point of the dynamics in Equation 6.17:

$$A_3^+ = \frac{\tau_- \tilde{A}_2^-}{\rho_{\text{targ}} \tau_+ \tau_y} \tag{6.19}$$

The postsynaptic rate should therefore roughly remain equal to its starting value ρ_{targ}. In practice, the Poisson assumption is not valid because of adaptation and refractoriness, and independence becomes violated as learning operates. This causes the postsynaptic firing rate to deviate and stabilize slightly away from the target ρ_{targ}. We therefore always set ρ_{targ} empirically so that the firing rate stabilizes to the true desired target.

6.2.4.3 Pair-based learning rule

We use a pair-based STDP rule structurally similar to the triplet rule described by Equation 6.16 (Figure 6.3B). The mechanism for LTD is identical, but LTP does not take into

account previous postsynaptic firing:

$$\frac{dw_j}{dt} = \eta_2 \left[A_2^+ r_j(t) y(t) - A_2^- o_1(t) x_j(t) \right] \tag{6.20}$$

where η_2 is the learning rate. A_2^- also slides in an activity-dependent manner according to Equation 6.18, to help stabilizing the output firing rate at a target ρ_{targ}. A_2^+ is set such that LTD initially balances LTP, i.e.

$$A_2^+ = \frac{\tilde{A}_2^- \tau_-}{\tau_+} \tag{6.21}$$

Comparing learning rules in a fair way requires making sure that their learning rates are equivalent. Since the two rules share the same LTD mechanism, we can simply take the same value for $\tilde{A}_2^-$ as well as $\eta_2 = \eta_3$. Since LTD is dynamically regulated to balance LTP on average in both rules, this ensures that they also share the same LTP rate.

6.2.4.4 Weight bounds

In order to prevent the weights from becoming negative or from growing too large, we set hard bounds on the synaptic efficacies for all three learning rules, when not stated otherwise. That is, if the learning rule requires a weight change Δw_j, w_j is set to

$$w_j \leftarrow \min \left[w_{\max}, \max \left(0, w_j + \Delta w_j \right) \right] \tag{6.22}$$

This type of bounds, in which the weight change is independent of the initial synaptic weight itself, is known to yield bimodal distributions of synaptic efficacies. In the simulation of Figure 6.5, we also consider the following soft bounds to extend the validity of our results to unimodal distributions of weights:

$$\begin{aligned} &\text{if } \Delta w_j \geq 0 \quad \text{then} \quad w_j \leftarrow w_j + \Delta w_j \\ &\text{if } \Delta w_j < 0 \quad \text{then} \quad w_j \leftarrow w_j + \left[1 - \frac{1}{1 + a \frac{w_j}{w_0}} + \left(\frac{1}{1+a} \right) \frac{w_j}{w_0} \right] \Delta w_j \end{aligned} \tag{6.23}$$

where a is a free parameter and $w_0 = 1$ mV is the value at which synaptic weights are initialized at the beginning of all learning experiments. This choice of soft-bounds is further motivated in the Results section. The shapes of the LTP and LTD weight-dependent factors are drawn in Figure 6.5A, for $a = 9$. Note that the LTD and LTP factors cross at w_0, which ensures that the balance between LTP and LTD set by Equation 6.19 and Equation 6.21 is initially preserved.

When the soft-bounds are used, the parameter τ_C of the optimal model is adjusted so that the weight distribution obtained with the optimal rule best matches the weight distributions

of the pair and triplet rules. This parameter indeed has an impact on the spread of the weight distribution: the optimal model knows about the generative model that underlies postsynaptic spike generation, and therefore takes optimally the noise into account, as long as τ_C spans no more than the width of the postsynaptic autocorrelation (Toyoizumi et al., 2005). If τ_C is equal to this width (about 20 ms), some weights can grow very large (¿50mV), which results in non-realistic weight distributions. Increasing τ_C imposes more detrimental noise such that all weights are kept within reasonable bounds. In order to constrain τ_C in a non-arbitrary way, we ran the learning experiment for several values of τ_C and computed the KL divergences between weight distributions (optimal-triplet, optimal-pair). τ_C is chosen to minimize these, as shown in Figure 6.5B.

6.2.5 Simulation of in vitro experiments

To obtain the predictions of the optimal model on standard *in vitro* STDP experiments, we compute the weight change of a single synapse ($N = 1$) according to Equation 6.12. The effect of the remaining thousands of synapses is concentrated in a large background noise, obtained by adding a $u_b = 19$ mV baseline to the voltage. The gain becomes $g_b = g(u_b) \simeq 21.45$ Hz, which in combination with adaptation and refractoriness would yield a spontaneous firing rate of about 7.5 Hz (see Figure 6.1). Spontaneous firing is artificially blocked, however. Instead, the neuron is forced to fire at precise times as described below.

The standard pairing protocol is made of a series of pre-post spike pairs, the spikes within the same pair being separated by $\Delta s = t_{\text{post}} - t_{\text{pre}}$. Pairs are repeated with some frequency f. The average $\bar{g}$ is taken fixed and equal to g_b, considering that STDP is optimal for *in vivo* conditions such that $\bar{g}$ should not adapt to the statistics of *in vitro* conditions. The homeostasis is turned off ($\gamma = 0$) in order to consider only the effects of the infomax principle.

6.3 Results

We study information transmission through a neuron modelled as a noisy communication channel. It receives input spike trains from a hundred plastic excitatory synapses, and stochastically generates output spikes according to an instantaneous firing rate modulated by presynaptic activities. Importantly, the firing rate is also modulated by the neuron's own firing history, in a way that captures the spike-frequency adaptation (SFA) mechanism found in a large number of cortical cell types. We investigate the ability of three different learning rules to enhance information transmission in this framework. The first learning rule is the standard pair-based STDP model, whereby every single pre-before-post (resp. post-before-pre)

Neuron model		**Optimal rule**		**Triplet rule**		**Pair rule**		**Weight bounds**	
τ_m	20 ms	η_o	0.04	η_3	1.0	η_2	1.0	$w_{\min}$	0 mV
g_0	1 Hz **(35)**	τ_C	20 ms	τ_+	16.8 ms	τ_+	16.8 ms	$w_{\max}$	4 mV
r_0	9.25 Hz *(3.25)*	τ_g	10 s	τ_-	33.7 ms	τ_-	33.7 ms	a	9
β	0.5 mV^{-1}	γ	1 **(0)**	τ_y	114 ms				
u_T	15 mV	g_{targ}	ad hoc	$\tilde{A}_2^-$	2.8e-3	$\tilde{A}_2^-$	2.8e-3		
τ_R	2 ms	λ	0.0094	A_3^+	6.5e-3	A_2^+	5.6e-3		
τ_A	150 ms			ρ_{targ}	ad hoc	ρ_{targ}	ad hoc		
q_R	100			τ_ρ	10 s	τ_ρ	10 s		
q_A	1 *(0)*								

Table 6.1: Baseline values of all parameters defined in the text. Some parameters were set to different values when the neuron was non-adapting (italic numbers). Similarly, some parameters were different for the simulations of *in vitro* experiment (bold faces)

spike pair yields LTP (resp. LTD) according to a standard double exponential asymmetric window (Bi and Poo, 1998; Song et al., 2000). The second one was developed in Pfister and Gerstner (2006) and is based on triplets of spikes. LTD is obtained similarly to the pair rule, whereas LTP is obtained from pairing a presynaptic spike with two postsynaptic spikes. The third learning rule (Toyoizumi et al., 2005) is derived from the infomax principle, under some metabolic constraints.

6.3.1 Triplet-STDP is better than pair-STDP when the neuron adapts

We assess and compare the performance of each learning rule on a simple spatiotemporal receptive field development task, with $N = 100$ presynaptic neurons converging onto a single postsynaptic cell (Figure 6.2A).

For each presynaptic neuron, a 5-second input spike train is generated once and for all (see Material and Methods). All presynaptic spike trains are then replayed continuously 5,000 times. All synapses undergo STDP according to one of the three learning rules. Synaptic weights are all initially set to 1 mV, which yields an initial output firing rate of about 7.5 Hz. We set the target firing rate ρ_{targ} of each learning rule such that the output firing rate stays very close to 7.5 Hz. To gather enough statistics, the whole experiment is repeated 10 times independently, each time with different input patterns. All results are therefore reported as mean and standard error of the mean (SEM) over the 10 trials.

All three learning rules developped very similar bimodal distributions of synaptic efficacies (Figure 6.4A), irrespective of the presence or absence of SFA. This is a well known consequence of additive STDP with hard bounds imposed on the synaptic weights (Kempter et al., 1999; Song et al., 2000). The firing rate stabilizes at 7.5 Hz as desired, for all plasticity rules

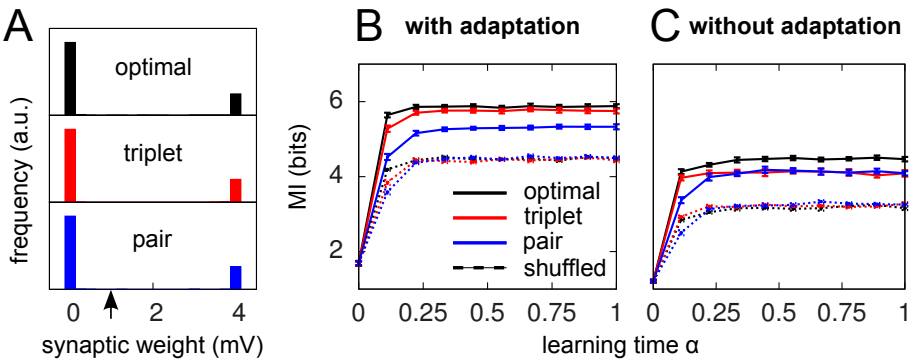


Figure 6.4: **Triplets are better than pairs when the neuron adapts. (A)** Distributions of synaptic efficacies obtained after learning. The weights were all initialized at 1 mV before learning (black arrow). When SFA is switched off, the very same bimodal distributions emerge (not shown). **(B)** Evolution of the MI along learning time. Learning time is arbitrarily indexed from $0 < \alpha < 1$. The dashed curves represent the MI when the weights taken from the momentary distribution at time α are shuffled. Each point is obtained from averaging the MI over 10 different shuffled versions of the synaptic weights. Error bars denote standard error of the mean (SEM) over 10 independent learning episodes with different input spike trains. **(C)** Same as in D, but SFA is switched off. The y-scale is the same as in **(B)**. Parameters for those simulations were $\lambda = 0$, $\gamma = 0$ with SFA, and $\gamma = 1$ without SFA. Other parameters took the values given in Table 6.1.

(not shown). In Figure 6.4B, we show the evolution of the MI (solid lines) as a function of learning time. It is computed as described in the Materials and Methods section, from the postsynaptic activity gathered during 100 periods (500 seconds). Since we are interested in quantifying the ability of different learning rules to enhance information transmission, we look at the information gain (defined as $\mathrm{MI}(\alpha = 1) - \mathrm{MI}(\alpha = 0)$) rather than the absolute value of the MI after learning. The triplet model reaches 98% of the "optimal" information gain while the pair model reaches 86% of it. Note that we call "optimal" what comes from the optimality model, but it is not necessarily the optimum in the space of solutions, because i) a stochastic gradient ascent may not always lead to the global maximum, ii) Toyoizumi et al's optimal learning rule involves a couple of approximations that may result in a sub-optimal algorithm (Toyoizumi et al., 2005), and iii) their learning rule does not specifically optimize information transmission for our periodic input scenario, but rather in a more general setting where input spike trains are drawn continuously from a fixed distribution (stationarity).

It is instructive to compare how much information is lost for each learning rule when the synaptic weights are shuffled. Shuffling means that the distribution stays exactly the same, while the detailed assignment of each w_j is randomized. The dashed lines in Figure 6.4B depict the MI under these shuffling conditions. Each point is obtained from averaging the MI over 10 different shuffled versions of the weights. The optimal and triplet model lose respectively 33% and 32% of their information gains, while the pair model loses only 23%.

This means that the optimal and triplet learning rules make a better choice in terms of the detailed assignment of each synaptic weight. For the pair learning rule, a larger part of the information gain is a mere side-effect of the weight distribution becoming bimodal. As an aside, we observe that the MI is the same (4.5 bits) in the "shuffled" condition for all three learning rules. This is an indication that we can trust our information comparisons. The result is also compatible with the value found by randomly setting 20 weights to the maximum value and the others to zero (Figure 6.2B, square mark).

How is adaptation involved in this increased channel capacity? In Figure 6.2C, the MI is plotted as a function of the postsynaptic firing rate, for an adaptive (black dots) and a non-adaptive (gray dots) neuron, irrespective of synaptic plasticity. Each point in the figure is obtained by setting randomly a given fraction χ of synaptic weights to the upper bound (4 mV), and the rest to 0 mV. The weight distribution stays bimodal, which leaves the neuron in a high information transmission state. χ is varied in order to cover a wide range of firing rates. We see that adaptation enhances information transmission at low firing rates (¡10Hz). The MI has a maximum at 7.5 Hz when the neuron is adapting (black circles). If adaptation is removed, the peak broadens and shifts to about 15 Hz (green circles). If the energetic cost of firing spikes is also taken into account, the best performance is achieved at 3 Hz, whether adaptation is enabled or not. This is illustrated in Figure 6.2C (lower plot) where the information per spike is reported as a function of the firing rate.

Is adaptation is beneficial in a general sense only, or does it differentially affect the three learning rules? To answer this question, we have the neuron learn again from the beginning, SFA being switched off. The temporal evolution of the MI for each learning rule is shown in Figure 6.4C. Overall, the MI is lower when the neuron does not adapt (compare panels B and C in Figure 6.4), which is in agreement with the previous paragraph and Figure 6.2C. Importantly, the triplet model loses its advantage over the pair model when adaptation is removed (compared red and blue lines in Figure 6.4C). This suggests a specific interaction between synaptic plasticity and the intrinsic postsynaptic dynamics in the optimal and triplet models. This is further investigated in later sections.

Finally, the main results of Figure 6.4 also hold when the distribution of weights remains unimodal. To achieve unimodal distributions with STDP, the hypothesis of hard-bounded synaptic efficacies must be relaxed. We implemented a form of weight-dependence of the weight change, such that LTP stays independent of the synaptic efficacy, while stronger synapses are depressed more strongly (see Methods). The weight-dependent factor for LTD had traditionally been modelled as being directly proportional to w_j (e.g. van Rossum et al. (2000)), which provides a good fit to the data obtained from cultured hippocampal neurons by Bi and Poo (1998). Morrison et al. (2007) proposed an alternative fit of the same data

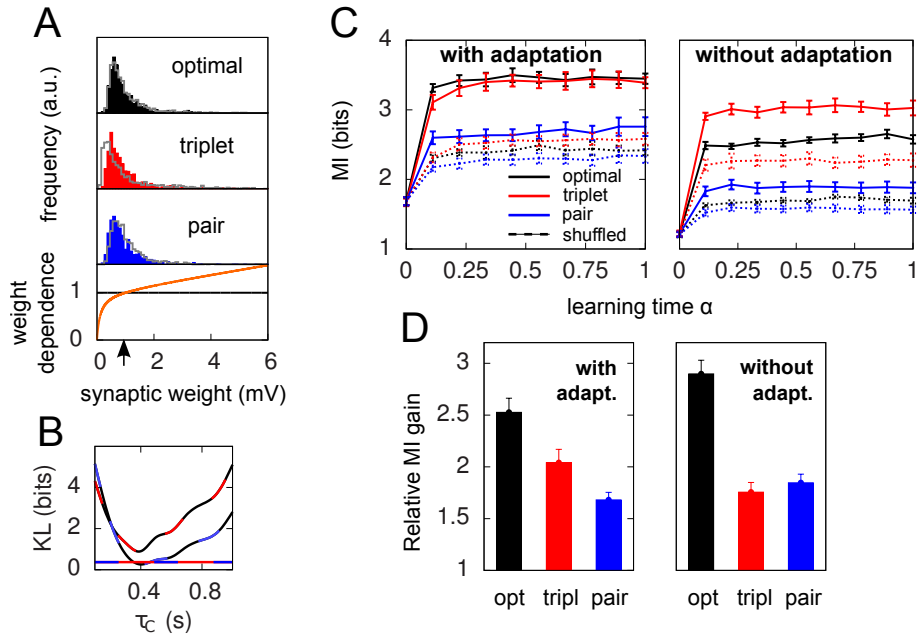


Figure 6.5: **Results hold for "soft-bounded" STDP.** The experiments of Figure 6.4 are repeated with soft-bounds on the synaptic weights (see Material and Methods). **(A)** Bottom: LTP is weight-independent (black line), whereas the amount of LTD required by each learning rule ($\Delta w < 0$) is modulated by a growing function of the momentary weight value (orange curve). The LTP and LTD curves cross at $w_0 = 1$ mV, which is also the initial value of the weights in our simulations. Top: this form of weight dependence produces unimodal but skewed distributions of synaptic weights after learning, for all three learning rules. The learning paradigm is the same as in Figure 6.4. Gray lines denote the weight distributions when adaptation is switched off. Note that histograms are computed by binning all weight values from all learning experiments, but the distributions look similar on individual experiments. In these simulations $\lambda = 0$, $a = 9$, and $\tau_C = 0.4$ s. **(B)** The parameter τ_C of the optimal learning rule has been chosen such that the weight distribution after learning stays as close as possible to that of the pair and triplet models. $\tau_C = 0.4$s minimizes the KL divergences between the distribution obtained from the optimal model and those from the pair (black-blue) and triplet (black-red) learning rules. The distance is then nearly as small as the triplet-pair distance (red-blue). **(C)** MI along learning time in this weight-dependent STDP scenario (cf. Figure 6.4B-C). **(D)** Normalized information gain (see text for definition).

with a different form of weight-dependence of LTP. Here we use a further alternative (see Methods, and Figure 6.5A). We require that the multiplicative factors for LTP and LTD exactly match at $w_j = w_0 = 1$ mV, where initial weights are set in our simulations. Further, we found it necessary that the slope of the LTD modulation around w_0 be less than one. Indeed, our neuron model is very noisy, such that reproducible pre-post pairs that need to be reinforced actually occur among a sea of random pre-post and post-pre pairs. If LTD too rapidly overcomes LTP above w_0, there is no chance for the correlated pre-post spikes to evoke sustainable LTP. The slope must be small enough for correlations to be picked up. This motivates our choice of weight dependence for LTD as depicted in Figure 6.5A. The weight

distributions for all three learning rules stay indeed unimodal, but highly positively skewed, such that the neuron can really "learn" by giving some relevant synapses large weights (tails of the distributions in Figure 6.5A). Note that the obtained weight distributions look like those recorded by Sjöström et al. (2001) (see e.g. Figure 3C in their paper).

The evolution of the MI along learning time is reported in Figure 6.5C. Overall, MI values are lower than those of Figure 6.4B. Unimodal distributions of synaptic efficaces are less informative than purely bimodal distributions, reflecting the lower degree of specialization to input features. Such distributions may however be advantageous in a memory storage task where old memories which are not recalled often need to be erased to store new ones. In this scenario, strong weights which become irrelevant can quickly be sent back from the tail to the main weight pool around 1mV. For a detailed study of the impact of the weight-dependence on memory retention, see Billings and van Rossum (2009).

We see that it is difficult to directly compare absolute values of the MI in Figure 6.5C, since the "shuffled" MIs (dashed lines) do not converge to the same value. This is because some weight distributions are more skewed than others (compare red and blue distributions in Figure 6.5A). In the present study, we are more interested in knowing how good our plasticity rules are at selecting individual weights for up- or down-regulation, on the basis of the input structure. We would like our performance measure to be free of the actual weight distribution, which is mainly shaped by the weight-dependence of Equation 6.23. We therefore compare the normalized information gain, i.e. $\frac{\mathrm{MI}(\alpha=1)-\mathrm{MI}(\alpha=0)}{\mathrm{MI}_{\mathrm{sh}}(\alpha=1)-\mathrm{MI}(\alpha=0)}$, where $\mathrm{MI}_{\mathrm{sh}}$ denotes the MI for shuffled weights. The result is shown in Figure 6.5D: the triplet is again better than the pair model, provided the postsynaptic neuron adapts.

Our simulations show that when SFA modulates the postsynaptic firing rate, the triplet model yields a better gain in information transmission than pair-STDP does. When adaptation is removed, this advantage vanishes. There must be a specific interaction between triplet-STDP and adaptation that we now seek to unravel.

6.3.2 Triplet-STDP increases the response entropy when the neuron adapts

Information transmission improves if the neuron learns to produce more diverse spike trains ($H(\mathbf{Y}^{\mathrm{K}})$ increases), and if the neuron becomes more reliable ($H(\mathbf{Y}^{\mathrm{K}}|\mathbf{Œ})$ decreases). In Figure 6.6A we perform a differential analysis of both entropies, on the same data as presented in Figure 6.4 (i.e. for hard-bounded STDP). Whether the postsynaptic neuron adapts (top) or not (bottom), the noise entropy (right) is drastically reduced, and the triplet learning rule does so better than the pair model (compare red and blue). The differential impact of adaptation on the two models can only be seen in the behaviour of the response entropy $H(\mathbf{Y}^{\mathrm{K}})$

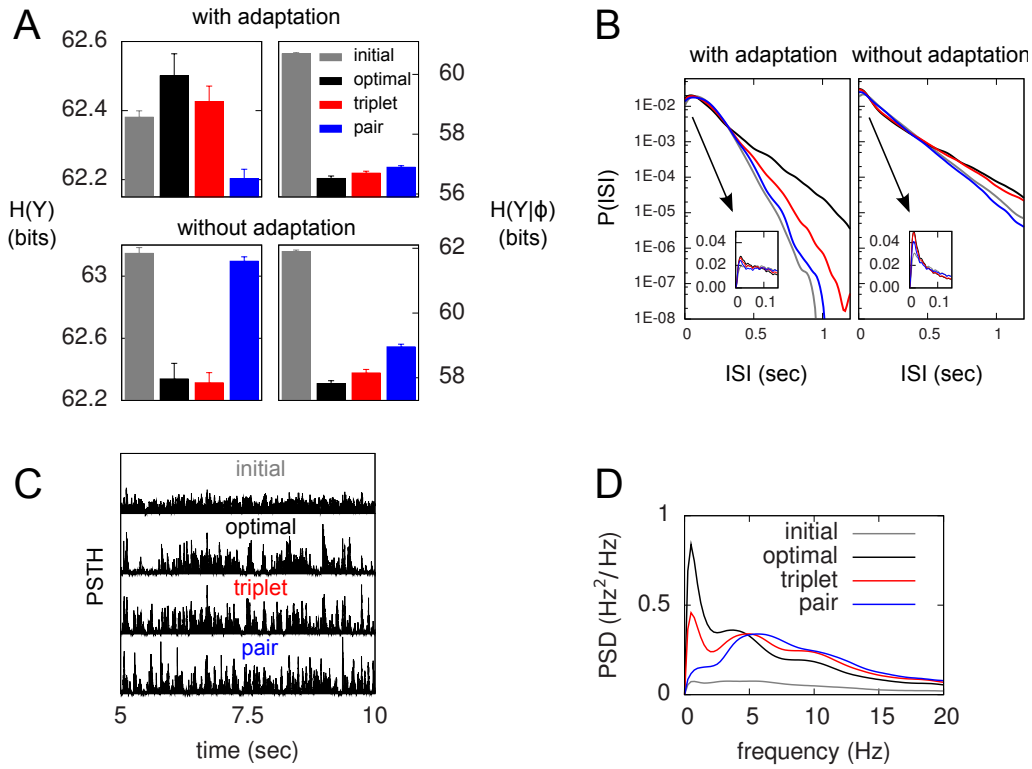


Figure 6.6: **Differential analysis of the entropies.** The learning experiments are the same as in Figure 6.4, using hard-bounds on the synaptic weights. **(A)** Response entropy (left) and noise entropy (right) with (top) and without (bottom) postsynaptic SFA. Entropies are calculated at the end of the learning process, except for the gray boxes which denote the entropies prior to learning. **(B)** Interspike-interval distributions with (left) and without (right) SFA, after learning (except gray line, before learning). The main plots have a logarithmic y-scale, whereas the insets have a linear one. **(C)** Peri-stimulus time histograms (PSTHs) prior to learning (top) and after learning for each learning rule, over a full 5-second period. All plots share the same y-scale. **(D)** Power spectra of the PSTHs shown in **(C)**, averaged over the 10 independent learning experiments.

(left). When the postsynaptic neuron adapts, triplet- and optimal-STDP both increase the response entropy, while it decreases with the pair model. This behaviour is reflected in the interspike-interval (ISI) distributions, shown in Figure 6.6B. With adaptation, the optimal and triplet rules produce distributions that are close to an exponential (which would be a straight line in the logarithmic y-scale). In contrast, the ISI distribution obtained from pair-STDP stays almost flat for ISIs between 25 and 120ms. Without adaptation, the optimal and triplet models further sparsifies the ISI distribution which then becomes sparser than an exponential, reducing the response entropy.

Qualitative similarities between the optimal and triplet models can also be found in the power spectrum of the Peri-Stimulus Time Histogram (PSTH). The PSTHs are plotted in Figure 6.6C over a full 5-second period, and their average power spectra are displayed in panel

D. The PSTH is almost flat prior to learning, reflecting the absence of feature selection in the input. Learning in all three learning rules creates sharp peaks in the PSTH, which illustrates the drop in noise entropy seen in panel A (right). The pair learning rule produces PSTHs with almost no power at low frequencies (below 5 Hz). In contrast, these low frequencies are strongly boosted by the optimal and triplet models. This is however not specific to SFA being on or off (not shown). We give an intuitive account for this in the Discussion.

This section has shed light on qualitative similarities in the way the optimal and triplet learning rules enhance information transmission in an adaptive neuron. We now seek to understand the reason why taking account of triplets of spikes would be close-to-optimal in the presence of postsynaptic SFA.

6.3.3 The optimal model exhibits a triplet effect

How similar is the optimal model to the triplet learning rule? In essence, the optimal model is a stochastic gradient learning rule, which updates the synaptic weights at every time step depending on the recent input-output correlations and the current relevance of the postsynaptic state. In contrast to this, phenomenological models require changing the synaptic efficacy upon spike occurrence only. It is difficult to compress what happens between spikes in the optimal model down to a single weight change at spike times. However we know that the dependence of LTP on previous postsynaptic firing is a hallmark of the triplet rule, and is absent in the pair rule. We therefore investigate the behavior of the optimal learning rule on post-pre-post triplets of spikes, and find a clear triplet effect (Figure 6.7).

We consider an isolated post-pre-post triplet of spikes, in this order (Figure 6.7A). Isolated means that the last pre- and postsynaptic spikes occured a very long time before this triplet. Let t^1_{post}, t_{pre} and t^2_{post} denote the spike times. The pre-post interval is kept constant equal to $\Delta s = t^2_{\text{post}} - t_{\text{pre}} = 15$ ms. We vary the length of the post-post interval $\Delta p = t^2_{\text{post}} - t^1_{\text{post}}$ from 16 ms to 500 ms. The resulting weight change is depicted in (Figure 6.7B). For comparison, the triplet model would produce – by construction – a decaying exponential with time constant τ_y. In the optimal model, potentiation decreases as the post-post interval increases. Two times constants show up in this decay, which reflect that of refractoriness (2 ms) and adaptation (150 ms). The same curve is drawn for two other adaptation time constants (see red and blue curves). When adaptation is removed, the triplet effect vanishes (dashed curve). It should be noted that the isolated pre-post pair itself (i.e. large post-post interval) results in a baseline amount of LTP, which is not the case in the triplet model. Figure 6.7A shows how this effect arises mechanistically. Three different triplets are shown, with the pre-post pair being fixed, and the post-post interval being either 16, 100, or 200

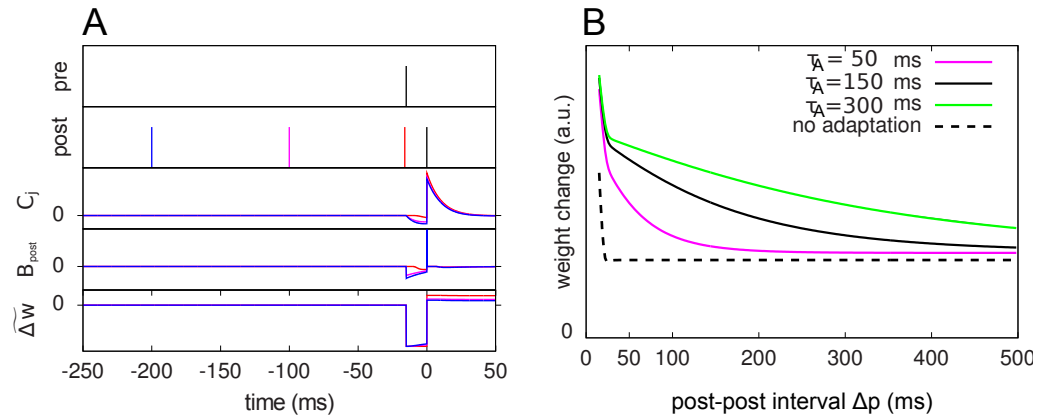


Figure 6.7: **The optimal model incorporates a triplet effect when the postsynaptic neuron adapts. (A)** A pre-post pair ($\Delta s = 15$ ms interval, black lines in the first two rows) is preceded by another postsynaptic spike. The post-post interval Δp is made either 16 ms (red line), 100 ms (purple) and 200 ms (blue). The time course of C_j, B_{post}, and the cumulative weight change $\widetilde{\Delta w}$ are plotted in the bottom rows. **(B)** Total weight change (optimal model) as a function of the post-post interval, for various adaptation time constants, and without adaptation (dashed line).

ms (red, purple, and blue respectively).

To further highlight the similarity between the optimal learning rule and the triplet model, we now derive an analytical expression for the optimal weight change that follows a post-pre-post triplet of spikes. Let us observe that the final cumulated weight change evoked by the triplet is dominated by the jump that occurs just following the second postsynaptic spike (Figure 6.7A) – except for the negative jump of size λ that follows the presynaptic spike arrival, but this is a constant. Our analysis therefore concentrates on the values of $C_j(t^2_{\text{post}})$ and $B_{\text{post}}(t^2_{\text{post}})$. Let us denote by $\epsilon_j = \exp\left(-\frac{\Delta s}{\tau_m}\right)$ the value of the unitary synaptic PSP at time t^2_{post}. Around the baseline potential $u_b = 19$ mV, the gain function is approximately linear (cf. Figure 6.1A), i.e. $g(u_b + w_j\epsilon_j) \simeq g_b + g'_b w_j \epsilon_j$ where $g_b = g(u_b)$ and $g'_b = \left.\frac{dg}{du}\right|_{u_b}$ are constants. From Equation 6.14, we read $B_{\text{post}}(t^2_{\text{post}}) = \log \frac{g_b + g'_b w_j \epsilon_j}{g_b} \delta(0)$, which is approximately equal to

$$B_{\text{post}}(t^2_{\text{post}}) \simeq \frac{g'_b}{g_b} w_j \epsilon_j \delta(0) \tag{6.24}$$

assuming the contribution of $w_j\epsilon_j$ is small compared to the baseline gain g_b. The term proportional to M in Equation 6.14 is negligible compared to the δ-function. From Equation 6.13, we see that

$$C_j(t^2_{\text{post}}) = \frac{\epsilon_j g'_b}{g_b + g'_b w_j \epsilon_j} + C_j(t^2_{\text{post}} - \varepsilon) \tag{6.25}$$

The total weight change following the second postsynaptic spike is therefore

$$\Delta w_j(t^2_{\text{post}}) \simeq \left(\frac{g'_b}{g_b}\right)^2 w_j \epsilon_j^2 + \frac{g'_b}{g_b} w_j \epsilon_j C_j(t^2_{\text{post}} - \varepsilon) \tag{6.26}$$

where

$$C_j(t^2_{\text{post}} - \varepsilon) = -\int_{t_{\text{pre}}}^{t^2_{\text{post}}-\varepsilon} \exp\left(-\frac{t^2_{\text{post}} - t}{\tau_C}\right) \exp\left(-\frac{t - t_{\text{pre}}}{\tau_m}\right) g'_b M(t) dt \tag{6.27}$$

Since we have taken $\tau_C = \tau_m$, the first two exponentials collapse into ϵ_j. To carry out the integration, let us further simplify the adaptation model into $M(t) = 1 - \exp(-(t - t^1_{\text{post}})/\tau_A)$, assuming that $t_{\text{pre}} - t^1_{\text{post}} > 2$ ms so that the refractoriness has already vanished at the time of the presynaptic spike, while adaptation remains. It is also assumed that the triplet is isolated, so that we can neglect the cumulative effect of adaptation. Equation 6.27 becomes

$$C_j(t^2_{\text{post}} - \varepsilon) = -\Delta s - \tau_A \exp\left(-\frac{\Delta p}{\tau_A}\right)\left[\exp\left(\frac{\Delta s}{\tau_A}\right) - 1\right] \tag{6.28}$$

If $\Delta s \ll \tau_A$, the last term into square brackets is approximately $\Delta s/\tau_A$. If not, ϵ_j^2 becomes so small that the whole r.h.s of Equation 6.28 vanishes. To sum up, the total weight change following the second postsynaptic spike is given by

$$\Delta w_j(t^2_{\text{post}}) = \frac{g'^2_b}{g_b} w_j \epsilon_j^2 \left(\frac{1}{g_b} - \Delta s\right) + \frac{g'^2_b}{g_b} \Delta s w_j \epsilon_j^2 \exp\left(-\frac{\Delta p}{\tau_A}\right) \tag{6.29}$$

The first term on the r.h.s of Equation 6.29 is a pair term, i.e. a weight change that depends only on the pre-post interval Δs. We note that it is proportional to ϵ_j^2, meaning that the time constant of the causal part of the STDP learning window is half the membrane time constant. The second term exactly matches the triplet model, when $\tau_A = \tau_y$ and $\tau_+ = \frac{\tau_m}{2}$. Indeed, the triplet model would yield the following weight change:

$$\Delta w_j^{\text{triplet}}(t^2_{\text{post}}) \simeq A_3^+ \epsilon_j \exp\left(-\frac{\Delta p}{\tau_y}\right) \tag{6.30}$$

From this we conclude that the triplet effect, which primarily arose from phenomenological minimal modeling of experimental data, also emerges from an optimal learning rule when the postsynaptic neuron adapts. To understand in more intuitive terms how the triplet mechanism relates to optimal information transmission, let us consider the case where the postsynaptic neuron is fully deterministic. If so, the noise entropy is null, so that maximizing information transfer means producing output spike trains with maximum entropy. If the mean firing rate ρ_{targ} is a further constraint, output spike trains should be Poisson processes, which as a by-product would produce exponentially distributed inter-spike intervals (ISIs). If the neuron is endowed with refractory and adapting mechanisms, there is a natural tendency for short ISIs to appear rarely. Therefore, plasticity has to fight against adaptation and refractoriness to bind more and more stimulus features to short ISIs. The triplet effect is precisely what is needed to achieve this: if a presynaptic spike is found to be responsible for a short ISI, it should be reinforced more than if the ISI was longer. This issue is further developped in the Discussion section.

6.3.4 Optimal STDP is target-cell specific

The results of the previous sections suggest that STDP may optimally interact with adaptation to enhance the channel capacity. In principle, if STDP is optimized for information transmission, it cannot ignore the intrinsic dynamics of the postsynaptic cell which influences the mapping between input and output spikes. The cortex is known to exhibit a rich diversity of cell types, with the corresponding range of intrinsic dynamics, and in parallel, STDP is target-cell specific (Lu et al., 2007; Tzounopoulos et al., 2004). Within the optimality framework, we should therefore be able to predict this target-cell specificity of STDP by investigating the predictions of the optimal model in the context of *in vitro* pairing experiments. Predictions should be made for different types of postsynaptic neurons, and be compared to experimental data. The optimal learning rule was shown in Toyoizumi et al. (2007) to share some features with STDP. We here extend this work to a couple of additional features including the frequency dependence. We also apply it to another type of postsynaptic cell, an inhibitory fast-spiking interneuron, for which *in vitro* data exist.

Only one synapse is investigated, with unit weight $w_0 = 1$ mV before the start of the experiment. 60 pre-post pairs with given inter-spike time Δs are repeated in time with frequency f. The subsequent weight change given by Equation 6.12 is reported as a function of both parameters (Figure 6.8, A and B).

The optimal model features asymmetric timing windows at 1, 20 and 50 Hz pairing frequencies (Figure 6.8A). At 1 and 20 Hz, pre-before-post yields LTP and post-before-pre leads to LTD. At 50 Hz the whole curve is shifted upwards, resulting in LTP on both sides. The model qualitatively agrees with the experimental data reported in Sjöström et al. (2001), redrawn for comparison (Figure 6.8A, circles).

The frequency dependence experimentally found in Markram et al. (1997) and Sjöström et al. (2001) is also qualitatively reproduced (Figure 6.8B). Post-pre pairing ($\Delta s = -10$ ms, green curve) switches from LTD at low frequency to LTP at higher frequencies, which is consistent with the timing windows in Figure 6.8A. For pre-post pairing ($\Delta s = +10$ ms, blue curve), LTP also increases with the pairing frequency. We also found that when SFA was removed, it was impossible to have a good fit for both the time window and the frequency dependence (not shown).

To further elucidate the link between optimal STDP and the after-spike kernel ($g_R + g_A$ in Equation 6.5), we ask whether plasticity at excitatory synapses onto fast-spiking (FS) interneurons can be accounted for in the same principled manner. In general, the intrinsic dynamics of inhibitory interneurons are very different from that of principal cells in cortex. STDP at synapses onto those cells is also different from STDP at excitatory-to-excitatory

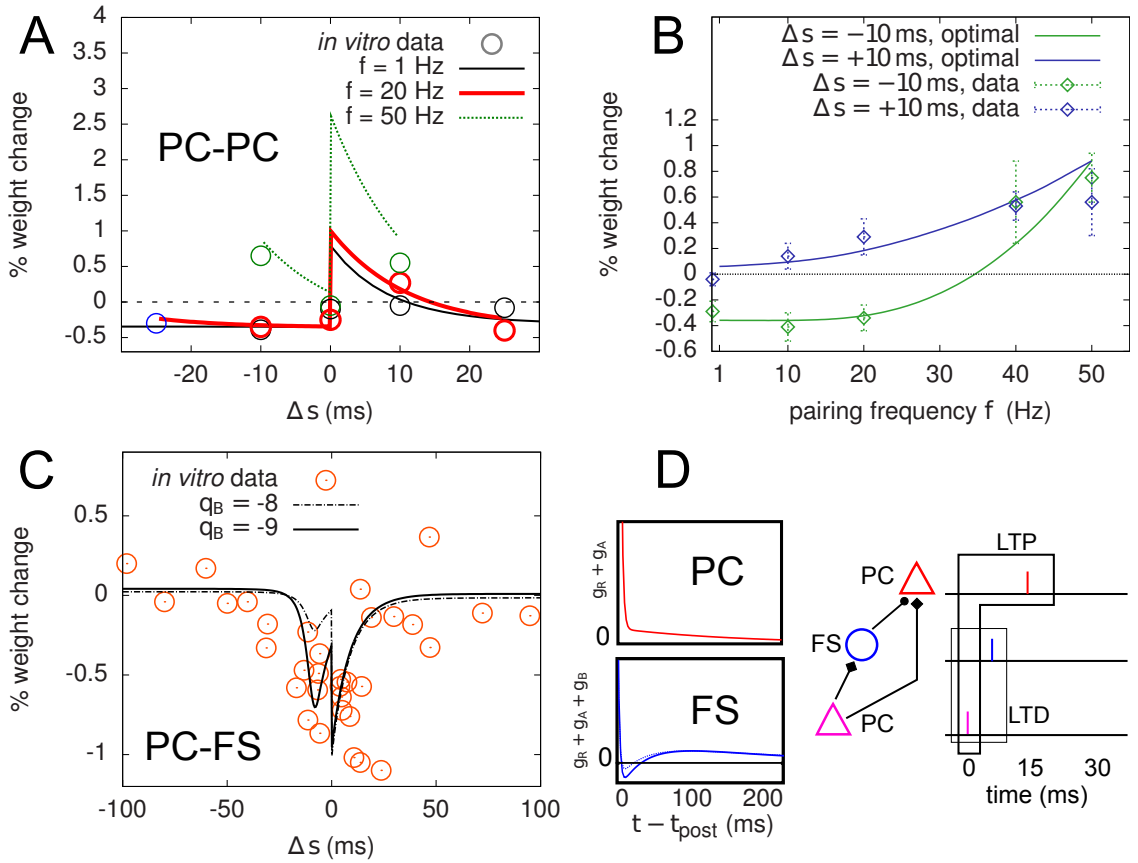


Figure 6.8: **Optimal plasticity shares features with target-cell specific STDP. (A)** The optimal model applied on 60 pre-post pairs repeating at 1 (black line), 20 (red thick) and 50 Hz (green) yields STDP learning windows that qualitatively match those recorded in Sjöström et al. (2001). For comparison, the *in vitro* data has been redrawn with permission. **(B)** LTP dominates when the pairing frequency is increased. The optimal frequency window is plotted for post-before-pre (−10 ms, solid green curve) and pre-before-post pairs (+10 ms, solid blue) repeated with frequency f (x-axis). Points and error bars are the experimental data, redrawn from Sjöström et al. (2001) with permission. **(C)** Learning window that minimizes information transmission at an excitatory synapse onto a fast-spiking (FS) inhibitory interneuron. The procedure is the same as in **(A)**. The spike-triggered adaptation kernel was updated to better match that of a FS cell (see **(D)**). Dots are redrawn from Lu et al. (2007). **(D)** Left: after-spike kernels of firing rate suppression for the principal excitatory cell (red, same as the one we used throughout the article, see Material and Methods) and the fast-spiking interneuron (blue). The latter was modeled by adding a third variable $q_B < 0$ with time constant $\tau_B = 30$ ms to the initial kernel. Solid blue line: $q_B = -9$. Dashed blue line: $q_B = -8$. Right: schematic of a feed-forward inhibition microcircuit. A first principal cell (PC) makes an excitatory connection to another PC. It also inhibits it indirectly through a FS interneuron. The example spike trains illustrate the benefit of having LTD for pre-before-post pairing at the PC-FS synapse (see text).

synapses (Lu et al., 2007; Tzounopoulos et al., 2004). The dynamics of FS cells are well modelled using a kernel which is shown in Figure 6.8D (Mensi et al., in preparation). We

augment the after-spike kernel with an additional variable g_B governed by

$$\frac{dg_B}{dt} = -\frac{g_B}{\tau_B} + q_B Y(t) \tag{6.31}$$

Parameters were set to $\tau_B = 30$ ms, $\tau_A = 150$ ms, $q_B = -9$ and $q_A = 4$. The resulting kernel (i.e. $g_R + g_A + g_B$ – Figure 6.8D, blue kernel) exhibits after-spike refractoriness followed by a short facilitating period before adaptation takes over (note that the kernel is suppressive, meaning that positive values correspond to suppression of activity while negative values mean facilitation). Since interneurons do not project over long distances to other areas, the infomax objective function might not appear as well justified. Instead, let us consider the simple microcircuit shown in Figure 6.8D. A first principal cell (PC) makes an excitatory synapse onto a second PC, and we assume the infomax principle is at work. The first PC inhibits the second PC via a fast-spiking (FS) interneuron. How, intuitively, should the PC-to-FS synapse change so that the FS cell also contributes to the overall information maximization between the two PCs? In a very crude understanding of the infomax principle, if a pre-before-post pair of spikes is evoked at the PC-PC synapse (see spike trains in Figure 6.8D), the probability of having this pair again should be increased. If a similar pre-before-post pair is simultaneously evoked at the PC-FS synapse, then decreasing its weight will make it less likely that the FS spike again after the first PC. This in turn makes it more likely that the first PC-PC pair of spike will occur again. Therefore, PC-FS synapses should undergo some sort of anti-Hebbian learning. In fact, we found information minimization (i.e. the optimal model with opposite learning rate) to yield a good match between the simulated STDP time window (Figure 6.8C) and that found in Lu et al. (2007), which also exhibits LTD on both sides with some LTP at large intervals (see orange dots, superimposed). The post-before-pre part of the window can be understood intuitively: when a presynaptic spike arrives a few milliseconds after a postsynaptic spike, it falls in the period where postsynaptic firing is facilitated ($q_B < 0$). Therefore, it still has some influence on the subsequent postsynaptic activity. In order to avoid later causal pre-post events, the weight should be decreased. We see that the optimal STDP window depends on the after-spike kernel that describes the dynamical properties of the postsynaptic cell: q_B directly modulates the post-pre part of the window (see dashed curve in Figure 6.8C).

Together, these results suggest that if STDP is considered as arising from an optimality principle, it naturally interacts with the dynamics of the postsynaptic cell. This might underlie the target-cell specificity of STDP (Lu et al., 2007; Tzounopoulos et al., 2004).

6.4 Discussion

Experiments (Markram et al., 1997; Sjöström et al., 2001; Froemke et al., 2006) as well as phenomenological models of STDP (Senn et al., 2001; Pfister and Gerstner, 2006; Froemke et al., 2006; Clopath et al., 2010) point to the fact that LTP is not accurately described by independent contributions from neighboring postsynaptic spikes. In order to reproduce the results of recent STDP experiments, at least two postsynaptic spikes must interact in the LTP process. We have shown that this key feature ("triplet effect" in Pfister and Gerstner (2006) and Clopath et al. (2010) and similarly in Senn et al. (2001)) happens to be optimal for an adapting neuron to learn to maximize information transmission. We have compared the performance of an optimal model (Toyoizumi et al., 2005) to that of two minimal STDP models. One of them incorporated the triplet effect (Pfister and Gerstner, 2006), while the second one did not (standard pair-based learning rule, Gerstner et al. (1996); Kempter et al. (1999); Song et al. (2000)). The triplet-based model performs very close to the optimal one, and this advantage over pair-STDP disappears when SFA is removed from the intrinsic dynamics of the postsynaptic cell.

Our results are not restricted to additive STDP in which the amount of weight change is independent of the weight itself. It also holds when the amount of LTD increases with the efficacy of the synapse, a form which better reflects experimental observations (Bi and Poo, 1998; Sjöström et al., 2001). In the model introduced here, the amount of LTD is modulated by a sub-linear function of the synaptic weight. The deviation from linearity is set by a single parameter $a > 0$, with the purely multiplicative dependence of van Rossum et al. (2000) being recovered when $a = 0$. Since we modeled only a fraction of the total input synapses, we assumed a certain level of noise in the postsynaptic cell to account for the activity of the remaining synapses, thereby staying consistent with the framework of information theory in which communication channels are generally considered noisy. Because of this noise level, we found a large a was required for the weight distribution to become positively skewed as reported by Sjöström et al. (2001) (cortex layer V). For both the pair and triplet learning rules, the noisier the postsynaptic neuron, the weaker the LTD weight-dependence (i.e. the larger a) must be to keep a significant spread of the weight distribution. This means that other (possibly simpler) forms of weight dependence for LTD would work equally well, provided the noise level is adjusted accordingly. For example, in a nearly deterministic neuron, input-output correlations are strong enough for the weight-distribution to spread even when LTD depends linearly on the synaptic weight ($a = 0$, not shown).

In the original papers where the optimal and triplet rule were first described, it was pointed out that both rules could be mapped onto the Bienenstock-Cooper-Munroe (BCM) learning rule (Bienenstock et al., 1982). Both learning rules are quadratic in the postsynaptic activity.

In turn, the link between the BCM rule and Independent-Components Analysis (ICA) has also already been researched (Intrator and Cooper, 1992; Blais et al., 1998; Clopath et al., 2010), as has the relationship between the infomax principle and ICA (Bell and Sejnowski, 1995). It therefore does not come as a surprise that the triplet model performs close to the infomax optimal learning rule. What is novel is the link to adaptation and spike after-potential.

We have also shown that when the optimal or triplet plasticity models are at work, the postsynaptic neuron learns to transmit information in a wider frequency band (Figure 6.6D): both rules evoke postsynaptic responses that have subtantial power below 5 Hz, in contrast to the pair-based STDP rule. This is intuitively understood from the triplet effect combined with adaptation. Let us imagine STDP starts creating a peak in the PSTH so that we have, with high probability, a first postsynaptic spike at time t_0. If a presynaptic spike at time $t_0 + \frac{\Delta}{2}$ is followed by a further postsynaptic spike at time $t_0 + \Delta$ (Δ on the order of 10ms), the triplet effect reinforces the connection from this presynaptic unit. In turn, it will create another peak at time $t_0 + \Delta$, and this process can continue. Peaks thus extend and become broader, until adaptation becomes strong enough to prevent further immediate firing. The next series of peaks will then be delayed by a few hundred milliseconds. Broadening of peak widths and inter-peak intervals together introduce more power at lower frequencies in the PSTH.

One should bear in mind that neurons process incoming signals in order to convey them to other receivers. Although the information content of the output spike train really is an important quantity with respect to information processing, the way it can be decoded by downstream neurons should also be taken into account. Some "words" in the output spike train may be more suited for subsequent transmission than others. It has been suggested (Lisman, 1997) that since cortical synapses are intrinsically unreliable, isolated incoming spikes cannot be received properly, whereas bursts of action potentials evoke a reliable response in the receiving neuron. There is a lot of evidence for burst firing in many sensory systems (see Krahe and Gabbiani (2004) for a review). As shown in Figure 6.6, the optimal and triplet STDP models tend to sparsify the distribution of inter-spike intervals, meaning that the neuron learns to respond vigorously (very short ISIs) to a larger number of features in the input stream, while remaining silent for longer portions of the stimulus. The neuron thus overcomes the effects of adaptation, which in baseline conditions (before learning) gives the ISI distribution a broad peak and a Gaussian-like drop-off. Our results therefore suggest that reliable occurence of short ISIs can arise from STDP in adaptive neurons that are not intrinsic bursters. This is in line with Eyherabide et al. (2008), which recently provided evidence for high information transmission through burst activity in an insect auditory system (*Locusta migratoria*). The recorded neurons encoded almost half of the total transmitted information in bursts, and this was also shown not to require intrinsic burst dynamics.

Since our results rely on the outcome of a couple of numerical experiments, one might be concerned about the validity of the findings outside the range of parameter values we have used. There are for example a couple of free parameters in the neuron model. It is obviously difficult to browse the full high-dimensional parameter space and search for regions where the results would break down. We therefore tried to constrain our neuron parameters in a sensible manner. For example, the parameters of the SFA mechanism (q_A and τ_A) were chosen such that the response properties to a step in input firing rate would look plausible (Figure 6.1C). The noise parameter r_0 and the threshold value u_T were chosen so as to achieve an output rate of 7.5Hz when all synaptic weights are at 1mV. We acknowledge, though, that r_0 could be made arbitrarily large (reducing the amount of noise) since u_T can compensate for it. In the limit of very low noise, information transmission cannot be improved by increasing the neuron's reliability anymore, since the noise entropy would already be minimal. We have shown however that a substantial part of the information gain found in the optimal and triplet models are due to an increased response entropy. This qualitative similarity, together with the structural similarities highlighted in Figure 6.7 and Figure 6.8, lead us to believe that our results would still hold in the deterministic limit, and for noise levels in between. The optimal plasticity rule becoming ill-defined in this limit, we did not investigate this further.

To what extent can we extrapolate our results to the optimality of synaptic plasticity in the real brain? It obviously depends on the amount of trust one can put into this triplet model. Phenomenological models of STDP are usually constructed based on the results of *in vitro* experiments. They end up reproducing the quantitative outcome of only a few pre-post pairing schemes which are far from spanning the full complexity of real spike trains. To what extent can these models be trusted in more natural situations? From a machine learning perspective, a minimal model is likely to generalize better than a more detailed model, because its small number of free parameters might prevent it from overfitting the experimental data at the expense of its interpolation/extrapolation power. In this study, we have put the emphasis on an extrapolation of recent minimal models (Pfister and Gerstner, 2006; Clopath et al., 2010): the amount of LTP obtained from a pre-before-post pair increases with the recent postsynaptic firing frequency. By construction, the models account for the frequency dependence of the classical pairing experiment (they are fitted on this, among other things). However, they are seriously challenged by a more detailed study of spike interactions at L2/3 pyramidal cells (Froemke et al., 2006). There, it was explicitly shown that (n-posts)-pre-post bursts yield an amount of LTD which grows with n, the number of postsynaptic spikes in the burst preceding the pair. In contrast, post-pre-post triplets in hippocampal slices lead to LTP in a way that is consistent with the triplet model (Wang et al., 2005). The results of our study should therefore be interpreted bearing in mind the variability in experimental results.

The recurrent *in vitro* versus *in vivo* debate should also be considered: synaptic plasticity depends on a lot of biochemical parameters for which the slice conditions do not faithfully reflect the normal operating mode of the brain.

A second controversy lies in our optimality model itself. While efficient coding of presynaptic spike trains may seem a reasonable goal to achieve at, say, thalamocortical synapses in sensory cortices, many other objectives could well be considered when it comes to other brain areas. Some examples are optimal decision making through risk balancing, reinforcement learning via reward maximization, or optimal memory storage and recall in autoassociative memories. It will be interesting to see more STDP learning rules in functionally different areas and how these relate to optimality principles.

Finally, while we investigated information transmission through a single postsynaptic cell, it remains to be elucidated how local information maximization in large recurrent networks of spiking neurons translates into a better information flow through the network.

Bibliography

Amit, D. J. and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, 7:237–252.

Aviel, Y. and Gerstner, W. (2006). From spiking neurons to rate models: A cascade model as an approximation to spiking neuron models with refractoriness. *Phys. Rev. E*, 73:051908.

Bartels, R. H. and Stewart, G. W. (1972). Solution of the matrix equation AX+XB=C. *Communications of the ACM*, 15:820–826.

Bell, A. J. and Parra, L. C. (2005). Maximising sensitivity in a spiking network. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Adv. Neural Inf. Process. Syst. 17*, pages 121–128. MIT Press, Cambridge, MA.

Bell, A. J. and Sejnowski, T. J. (1995). An information maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7:1129–1159.

Bell, C. C., V., H., Sugawara, Y., and Grant, K. (1997). Synaptic plasticity in a cerebellum-like structure depends on temporal order. *Nature*, 387:278–281.

Ben-Yishai, R., Bar-Or, R. L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. USA*, 92:3844–3848.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331:83–87.

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

Bi, G. Q. and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, 18:10464–10472.

Bienenstock, E. L., Cooper, L. N., and Munroe, P. W. (1982). Theory of the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2:32–48.

Billings, G. and van Rossum, M. C. W. (2009). Memory retention and spike-timing dependent plasticity. *J. Neurophysiol.*, 101:2775–2788.

Blais, B. S., Intrator, N., Shouval, H., and Cooper, L. N. (1998). Receptive field formation in natural scene environments: comparison of single-cell learning rules. *Neural Comput.*, 10:1797–1813.

Bohte, S. M. and Mozer, M. C. (2007). Reducing the variability of neural responses: A computational theory of spike-timing-dependent plasticity. *Neural Comput.*, 19:371–403.

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

Buonomano, D. V. and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.*, 10:113–125.

Burke, J. V., Lewis, A. S., and Overton, M. L. (2002). Two numerical methods for optimizing matrix stability. *Lin. Alg. and its Appl.*, 351:117145.

Cafaro, J. and Rieke, F. (2010). Noise correlations improve response fidelity and stimulus encoding. *Nature*, 468:964–967.

Chechik, G. (2003). Spike timing-dependent plasticity and relevant mututal information maximization. *Neural Comput.*, 15:1481–1510.

Chklovskii, D. B. (2004). Synaptic connectivity and neuronal morphology: two sides of the same coin. *Neuron*, 43:609–617.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487:51–56.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I., and Shenoy, K. V. (2010a). Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68:387400.

Churchland, M. M. and Shenoy, K. V. (2007). Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J. Neurophysiol.*, 97:4235–4257.

Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., Bradley, D. C., Smith, M. A., Kohn, A., Movshon, J. A., Armstrong, K. M., Moore, T., Chang, S. W., Snyder, L. H., Lisberger, S. G., Priebe, N. J., Finn, I. M., Ferster, D., Ryu, S. I., Santhanam, G., Sahani, M., and Shenoy, K. V. (2010b). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.*, 13:369–378.

Clopath, C., Büsing, L., Vasilaki, E., and Gerstner, W. (2010). Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nat. Neurosci.*, 13:344–352.

Clopath, C., Longtin, A., and Gerstner, W. (2008). An online hebbian learning rule that performs Independent Component Analysis. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Adv. Neural Inf. Process. Syst. 20*, pages 321–328. MIT Press, Cambridge, MA.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

de Franciscis, S., Johnson, S., and Torres, J. J. (2011). Enhancing neural-network performance via assortativity. *Phys. Rev. E*, 83:036114.

Dorrn, A. L., Yuan, K., Barker, A. J., Schreiner, C. E., and Froemke, R. C. (2010). Developmental sensory experience balances cortical excitation and inhibition. *Nature*, 465:932–936.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A., and Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, 269:981–985.

Douglas, R. J. and Martin, K. A. C. (2007). Recurrent neuronal circuits in the neocortex. *Cur. Biol.*, 17:R496–R500.

Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., and Apkarian, A. V. (2005). Scale-Free brain functional networks. *Phys. Rev. Lett.*, 94:018102.

Ermentrout, B. (1994). Reduction of conductance-based models with slow synapses to neural nets. *Neural Comput.*, 6:679695.

Eyherabide, H. G., Rokem, A., Herz, A. V. M., and I., S. (2008). Burst firing is a neural code in an insect auditory system. *Front. Comput. Neurosci.*, 2:1–17.

Ferezou, I., Haiss, F., Gentet, L. J., Aronoff, R., Weber, B., and Petersen, C. C. H. (2007). Spatiotemporal dynamics of cortical sensorimotor integration in behaving mice. *Neuron*, 56:907–923.

Fino, E. and Yuste, R. (2011). Dense inhibitory connectivity in neocortex. *Neuron*, 69:1188–1203.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cog. Sci.*, 14:119–130.

Fiser, J., Chiu, C., and Weliky, M. (2004). Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature*, 431:573–578.

Florian, R. V. (2007). Reinforcement learning through modulation of spike timing-dependent synaptic plasticity. *Neural Comput.*, 19:1468–1502.

Froemke, R. C., Merzenich, M. M., and Schreiner, C. E. (2007). A synaptic memory trace for cortical receptive field plasticity. *Nature*, 450:425–429.

Froemke, R. C., Tsay, I., Raad, M., Long, J., and Dan, Y. (2006). Contribution of individual spikes in burst-induced long-term synaptic modification. *J. Neurophysiol.*, 95:1620–1629.

Fuster, J. M. and Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, 173:652654.

Ganguli, S., Bisley, J. W., Roitman, J. D., Shadlen, M. N., Goldberg, M. E., and Miller, K. D. (2008a). One-dimensional dynamics of attention and decision making in LIP. *Neuron*, 58:15–25.

Ganguli, S., Huh, D., and Sompolinsky, H. (2008b). Memory traces in dynamical systems. *Proc. Natl. Acad. Sci. USA*, 105:18970–18975.

Ganguli, S. and Latham, P. (2009). Feedforward to the past: The relation between neuronal connectivity, amplification, and short-term memory. *Neuron*, 61:499–501.

Gardiner, C. W. (1985). *Handbook of stochastic methods: for physics, chemistry, and the natural sciences*. Berlin: Springer.

Gentet, L. J., Avermann, M., Matyas, F., Staiger, J. F., and Petersen, C. C. H. (2010). Membrane potential dynamics of GABAergic neurons in the barrel cortex of behaving mice. *Neuron*, 65:422–435.

Gerstner, W. (2000). Population dynamics of spiking neurons: fast transients, asynchronous states, and locking. *Neural Comput.*, 12:43–89.

Gerstner, W., Kempter, R., van Hemmen, J., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383:76–78.

Gerstner, W. and Kistler, W. K. (2002a). Mathematical formulations of Hebbian learning. *Biol. Cybern.*, 87:404–415.

Gerstner, W. and Kistler, W. M. (2002b). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge Univ. Pr.

Gerstner, W. and Naud, R. (2009). How good are neuron models? *Science*, 326:379–380.

Gilbert, C. D. and Wiesel, T. N. (1983). Clustered intrinsic connections in cat visual cortex. *J. Neurosci.*, 3:1116–1133.

Goh, K., Kahng, B., and Kim, D. (2001). Spectra and eigenvectors of scale-free networks. *Phys. Rev. E*, 64:051903.

Goldberg, J. A., Rokni, U., and Sompolinsky, H. (2004). Patterns of ongoing activity and the functional architecture of the primary visual cortex. *Neuron*, 42:489–500.

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron*, 61:621–634.

Grabow, C., Grosskinsky, S., and Timme, M. (2012). Small-World network spectra in mean-field theory. *Phys. Rev. Lett.*, 108:218701.

Hellwig, B. (2000). A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biol. Cybern.*, 82:111121.

Hennequin, G., Vogels, T. P., and Gerstner, W. (2012). Non-normal amplification in random balanced neuronal networks. *Phys. Rev. E*, 86:011909.

Intrator, N. and Cooper, L. (1992). Objective function formulation of the bcm theory of visual cortical plasticity – statistical connections, stability conditions. *Neural Netw.*, 5:3–17.

Izhikevich, E. and Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proc. Natl. Acad. Sci. U.S.A.*, 105:3593–3598.

Izhikevich, E. M. (2006). *Dynamical systems in neuroscience: the geometry of excitability and bursting*. MIT press.

Kempter, R., Gerstner, W., and van Hemmen, J. L. (1999). Hebbian learning and spiking neurons. *Phys. Rev. E*, 59:4498–4514.

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., and Arieli, A. (2003). Spontaneously emerging cortical representations of visual attributes. *Nature*, 425:954–956.

Klampfl, S., Legenstein, R., and Maass, W. (2009). Spiking neurons can learn to solve information bottleneck problems and to extract independent components. *Neural Comput.*, 21:911–959.

Krahe, R. and Gabbiani, F. (2004). Burst firing in sensory systems. *Nat. Rev. Neurosci.*, 5:13–23.

Kriener, B., Tetzlaff, T., Aertsen, A., Diesmann, M., and Rotter, S. (2008). Correlations and population dynamics in cortical networks. *Neural Comput.*, 20:2185–2226.

Kullmann, D. M., Moreau, A. W., Bakiri, Y., and Nicholson, E. (2012). Plasticity of inhibition. *Neuron*, 75:951–962.

Kumar, A., Schrader, S., Aertsen, A., and Rotter, S. (2008). The high-conductance state of cortical networks. *Neural Comput.*, 20:1–43.

Ledoux, E. and Brunel, N. (2011). Dynamics of networks of excitatory and inhibitory neurons in response to time-dependent inputs. *Front. Comput. Neurosci.*, 5:1–25.

Lengyel, M., Kwag, J., Paulsen, O., and Dayan, P. (2005). Matching storage and recall: hippocampal spike timing-dependent plasticity and phase response curves. *Nat. Neurosci.*, 8:1677–1683.

Lerchner, A., Sterner, G., Hertz, J., and Ahmadi, M. (2006). Mean field theory for a balanced hypercolumn model of orientation selectivity in primary visual cortex. *Network: Comput. Neural Sys.*, 17:131150.

Levy, R. B. and Reyes, A. D. (2012). Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *J. Neurosci.*, 32:5609–5619.

Linsker, R. (1989). How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Comput.*, 1:402–411.

Lisman, J. (1997). Bursts as a unit of neural information: making unreliable synapses reliable. *Trends in Neurosci.*, 20:38–43.

Lu, J., Li, C., Zhao, J.-P., Poo, M., and Zhang, X. (2007). Spike-timing-dependent plasticity of neocortical excitatory synapses on inhibitory interneurons depends on target cell type. *J. Neurosci.*, 27:9711–9720.

Luczak, A., Barthó, P., and Harris, K. D. (2009). Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron*, 62:413–425.

Luz, Y. and Shamir, M. (2012). Balancing feed-forward excitation and inhibition via hebbian inhibitory synaptic plasticity. *PLoS Comput. Biol.*, 8:e1002334.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.*, 14:25312560.

Magee, J. C. and Johnston, D. (1997). A synaptically controlled associative signal for hebbian plastiticy in hippocampal neurons. *Science*, 275:209–213.

Mariño, J., Schummers, J., Lyon, D. C., Schwabe, L., Beck, O., Wiesing, P., Obermayer, K., and Sur, M. (2005). Invariant computations in local cortical networks with balanced excitation and inhibition. *Nat. Neurosci.*, 8(2):194201.

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

Mehta, M. L. (2004). *Random matrices*. Academic press.

Miller, K. D. and Fumarola, F. (2011). Mathematical equivalence of two common forms of firing rate models of neural networks. *Neural Comput.*, 24:25–31.

Molgedey, L., Schuchhardt, J., and Schuster, H. G. (1992). Suppressing chaos in neural networks by noise. *Phys. Rev. Lett.*, 69:3717–3719.

Morrison, A., Aertsen, A., and Diesmann, M. (2007). Spike-timing dependent plasticity in balanced random networks. *Neural Comput.*, 19:1437–1467.

Morrison, A., Diesmann, M., and Gerstner, W. (2008). Phenomenological models of synaptic plasticity based on spike timing. *Biol. Cybern.*, 98:459–478.

Murphy, B. K. and Miller, K. D. (2009). Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61:635–648.

Noll, D. and Apkarian, P. (2005). Spectral bundle methods for non-convex maximum eigenvalue functions: second-order methods. *Math. progr.*, 104(2):729747.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, 15:267–273.

Oja, E. (1989). Neural networks, principal components, and subspaces. *Int. J. Neural Syst.*, 1:61–68.

Okun, M. and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci.*, 11:535–537.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Ostojic, S. and Brunel, N. (2011). From spiking neuron models to linear-nonlinear models. *PLoS Comput. Biol.*, 7(1):e1001056.

Ozeki, H., Finn, I. M., Schaffer, E. S., Miller, K. D., and Ferster, D. (2009). Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62:578–592.

Packer, A. M. and Yuste, R. (2011). Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: A canonical microcircuit for inhibition? *J. Neurosci.*, 31:13260–13271.

Perin, R., Berger, T. K., and Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci. USA*, 108:5419–5424.

Pernice, V., Staude, B., Cardanobile, S., and Rotter, S. (2011). How structure determines correlations in neuronal networks. *PLoS Comput. Biol.*, 7:e1002059.

Pfister, J.-P. and Gerstner, W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. *J. Neurosci.*, 26:9673–9682.

Pfister, J.-P., Toyoizumi, T., Barber, D., and Gerstner, W. (2006). Optimal Spike-Timing Dependent Plasticity for precise action potential firing in supervised learning. *Neural Comput.*, 18:1309–1339.

Poulet, J. F. A. and Petersen, C. C. H. (2008). Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice. *Nature*, 454:881–885.

Rajan, K. and Abbott, L. F. (2006). Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.*, 97:188104.

Rajan, K., Abbott, L. F., and Sompolinsky, H. (2010). Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, 011903:1–5.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, 2:79–87.

Renart, A., de la Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., and Harris, K. (2010). The asynchronous state in cortical circuits. *Science*, 327:587.

Richardson, M. (2009). Dynamics of populations and networks of neurons with voltage-activated and calcium-activated currents. *Phys. Rev. E*, 80(2).

Richardson, M. and Swarbrick, R. (2010). Firing-rate response of a neuron receiving excitatory and inhibitory synaptic shot noise. *Phys. Rev. Lett.*, 105.

Roxin, A. (2011). The role of degree distribution in shaping the dynamics in networks of sparsely connected spiking neurons. *Front. Comput. Neurosci.*, 5:1–15.

Rubin, J., Lee, D. D., and Sompolinsky, H. (2001). Equilibrium properties of temporally asymmetric Hebbian plasticity. *Phys. Rev. Lett.*, 86:364–367.

Senn, W., Tsodyks, M., and Markram, H. (2001). An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing. *Neural Comput.*, 13:35–67.

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40:1063–1073.

Shefi, O., Golding, I., Segev, R., Ben-Jacob, E., and Ayali, A. (2002). Morphological characterization of in vitro neuronal networks. *Phys. Rev. E*, 66:021905.

Shenoy, K. V., Kaufman, M. T., Sahani, M., and Churchland, M. M. (2011). A dynamical systems view of motor preparation: Implications for neural prosthetic system design. *Progr. Brain Res.*, 192:33.

Shriki, O., Hansel, D., and Sompolinsky, H. (2003). Rate models for conductance-based cortical neuronal networks. *Neural Comput.*, 15:1809–1841.

Sjöström, P., Turrigiano, G., and Nelson, S. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32:1149–1164.

Smith, E. C. and Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, 439:978–982.

Somers, D. C., Nelson, S. B., and Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.*, 15:54485465.

Sommers, H. J., Crisanti, A., Sompolinsky, H., and Stein, Y. (1988). Spectrum of large random asymmetric matrices. *Phys. Rev. Lett.*, 60:1895–1898.

Sompolinsky, H., Crisanti, A., and Sommers, H. J. (1988). Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262.

Sompolinsky, H. and Shapley, R. (1997). New perspectives on the mechanisms for orientation selectivity. *Curr. Op. Neurobiol.*, 7:514522.

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive Hebbian learning through spike-time-dependent synaptic plasticity. *Nat. Neurosci.*, 3:919–926.

Sporns, O. (2011). The Non-Random brain: efficiency, economy, and complex dynamics. *Front. Comput. Neurosci.*, 5.

Sprekeler, H., Hennequin, G., and Gerstner, W. (2009). Code-specific policy gradient rules for spiking neurons. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Adv. Neural Inf. Process. Syst. 22*, pages 1741–1749.

Sprekeler, H., Michaelis, C., and Wiskott, L. (2007). Slowness: An objective for spike-timing-plasticity? *PLoS Comp. Biol.*, 3:e112.

Sussillo, D. and Abbott, L. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557.

Tao, T. (2011). Outliers in the spectrum of iid matrices with bounded rank perturbations. *Proba. Th. Relat. Fields*, Online First:1–33. 10.1007/s00440-011-0397-9.

Toyoizumi, T., Pfister, J., Aihara, K., and Gerstner, W. (2007). Optimality Model of Unsupervised Spike-Timing-Dependent Plasticity: Synaptic Memory and Weight Distribution. *Neural Comput.*, 19:639.

Toyoizumi, T., Pfister, J.-P., Aihara, K., and Gerstner, W. (2005). Generalized bienenstock-cooper-munro rule for spiking neurons that maximizes information transmission. *Proc. Natl. Acad. Sci. U.S.A.*, 102:5239–5244.

Trefethen, L. N. and Embree, M. (2005). *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press.

Trefethen, L. N., Trefethen, A. E., Reddy, S. C., and Driscoll, T. A. (1993). Hydrodynamic stability without eigenvalues. *Science*, 261:578–584.

Tsodyks, M., Kenet, T., Grinvald, A., and Arieli, A. (1999). Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286:1943–1946.

Tsodyks, M. V., Skaggs, W. E., Sejnowski, T. J., and McNaughton, B. L. (1997). Paradoxical effects of external modulation of inhibitory interneurons. *J. Neurosci.*, 17:43824388.

Tzounopoulos, T., Kim, Y., Oertel, D., and Trussell, L. . (2004). Cell-specific, spike timing-dependent plasticity in the dorsal cochlear nucleus. *Nat. Neurosci.*, 7:719–725.

van Rossum, M. C. W., Bi, G. Q., and Turrigiano, G. G. (2000). Stable Hebbian learning from spike timing-dependent plasticity. *J. Neurosci.*, 20:8812–8821.

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

Vanbiervliet, J., Vandereycken, B., Michiels, W., Vandewalle, S., and Diehl, M. (2009). The smoothed spectral abscissa for robust stability optimization. *SIAM Journal on Optimization*, 20:156171.

Vogels, T. P. and Abbott, L. F. (2009). Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nat. Neurosci.*, 12:483–491.

Vogels, T. P., Rajan, K., and Abbott, L. F. (2005). Neural network dynamics. *Annu. Rev. Neurosci.*, 28:357–376.

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

Voges, N., Schüz, A., Aertsen, A., and Rotter, S. (2010). A modeler's view on the spatial structure of intrinsic horizontal connectivity in the neocortex. *Progr. Neurobiol.*, 92:277–292.

Wang, H.-X., Gerkin, R. C., Nauen, D. W., and Wang, G.-Q. (2005). Coactivation and timing-dependent integration of synaptic potentiation and depression. *Nat. Neurosci.*, 8:187–193.

Wang, X. (1999). Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.*, 19:9587–9603.

Wehr, M. and Zador, A. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, 426:442446.

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J*, 12:124.

Woodin, M. A., Ganguly, K., Poo, M., et al. (2003). Coincident pre-and postsynaptic activity modifies GABAergic synapses by postsynaptic changes in cl-transporter activity. *Neuron*, 39:807820.

Xie, X. and Seung, H. S. (2004). Learning in neural networks by reinforcement of irregular spiking. *Phys. Rev. E*, 69:041909.

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A., and Poo, M. M. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395:37–44.

Curriculum Vitae

Guillaume Hennequin, born Sept. 28th, 1985. Married, one child.

■ Academic training

10.2007 – present	PhD studies in computational neuroscience, Wulfram Gerstner's lab. Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
10.2006 – 11.2007	Master of Science in Cognitive Science, the University of Edinburgh (UK). Awarded with distinction.
09.2004 – 11.2007	Ecole Supérieure d'Eléctricité (SUPELEC, top French engineering school). Undergraduate studies in electrical engineering and computer science. Final year in Edinburgh (cf. above).
09.2002 – 06.2004	French "classes préparatoires" : intensive training in mathematics, physics and chemistry, for competitive entry in top engineering schools.

■ Additional training and experience

10.2011	One-week workshop on mean-field methods applied to neuroscience (Marseille, France).
10.2010	2-day workshop on spike-frequency adaptation (Dresden, Germany). I had a poster to explain my work on synaptic plasticity and how it relates to spike-frequency adaptation and other forms of intrinsic neuronal dynamics in the context of information maximisation.
08.2008	3-week summer school in theoretical neuroscience, organised by the Frankfurt Institute of Advanced Studies (FIAS). Miniproject on modeling adult receptive field plasticity in the auditory cortex.
07.2006 – 08.2006	Research internship at LORIA, "Cortex" team (`http://cortex.loria.fr`).

Publications

- Amplification and rotational dynamics in inhibition-stabilized cortical circuits
  G. Hennequin, T. P. Vogels and W. Gerstner
  close to submission – chapter 3 of this thesis
- Non-normal amplification in random neuronal networks
  G. Hennequin, T. P. Vogels and W. Gerstner
  Physical Review E (2012) – chapter 2 of this thesis
- STDP in adaptive neurons gives close-to-optimal information transmission
  G. Hennequin, W. Gerstner and J.-P. Pfister
  Frontiers in Computational Neuroscience (2010) – chapter 6 of this thesis
  (**14 citations so far – Google Scholar**)
- Code-specific policy gradient learning rules for spiking neurons
  H. Sprekeler, **G. Hennequin** and W. Gerstner
  Neural Information and Processing Systems (2009)
- Computational explorations in perceptual learning
  G. Hennequin
  MSc dissertation, University of Edinburgh (2007) – this work received a formal "distinction"

Conference abstracts

2012 **Hennequin G**, Vogels TP, Gerstner W
Nonnormal amplification in random balanced networks
Cosyne abstract, Salt Lake City (February)

2011 **Hennequin G**, Vogels TP, Gerstner W
Fast and richly structured activity in cortical networks with local inhibition
CNS abstract, Stockholm (July)

Zenke F, **Hennequin G**, Sprekeler H, Vogels TP and Gerstner W
Plasticity and stability in recurrent neural networks
CNS abstract, Stockholm (July)

2009 **Hennequin G**, Pfister J-P, Gerstner W
STDP interacts with neural dynamics to enhance information transmission
Cosyne abstract, Salt Lake City (February)

■ Selected presentations

January 2012 — Transient amplification in the cortex: what are "good" connectivity motifs?
Neurotheory Center
Columbia University, NY, USA

January 2012 — Amplification and fast dynamics in cortical circuits: the role of synaptic connectivity
Computational and Biological Learning Laboratory
The University of Cambridge, UK (postdoctoral position interview)

■ Referee activities

I regularly review papers submitted to the following journals: PLoS Computational Biology, PLoS One, Neural Computation, Frontiers in Computational Neuroscience.

■ Grants and awards

Following my job interview for a PhD position at EPFL, I was **awarded a one-year Excellence Scholarship** from the Lemanic Neuroscience program (10.2007 – 10.2008).