Universität Bremen

Dissertation in partial fulfilment of the degree
Doctor rerum naturalium (Dr. rer. nat.)

The role of long-range connections in contextual processing and spontaneous activity of primary visual cortex

Submitted by
Federica Capparelli
May 26th, 2020

Gutachter
Prof. Dr. Klaus Pawelzik
Prof. Dr. Andreas Kreiter

I would like to dedicate a thought to all the colleagues and friends that contributed to this work, with clever suggestions, annoying suggestions, cheerful comments, patient and submissive reading or even just a phone call from far away. I don't do well with *etiquettes*, so forgive me if I don't follow one now.

Udo, David, Klaus, I need and want to thank each of you for your supervision. For the guidance, the technical support, the things you let me try, fail and repeat until they became 'perfect'. For the education I received, the science I learned and the trust you treated me with. When I moved to Bremen I was already an adult, I don't know how to define the degree of maturity and awareness that I reached today.

I also want to thank all the members of the committee, for agreeing to virtually join my colloquium in this special and uncertain circumstances. Especially Andreas Kreiter, for taking the time to read and evaluate my thesis and Michael Mackey, for reading the earliest version of this work and for his kind and unexpected words of encouragement.

Nergis and Dmitriy, you took great care of me, much more than you would ever realize. Listening to your stories and to your animated conversations had the power to chase away loneliness and, without you here, it's been a little harder. I did and will always look up to you for the passionate way you live your life.
Daniel, Axel, Eric, Serge, another Erik, but with a different spelling, Maik – yes, I'm going in order of seniority, or sort of, just to make someone mad – Alina, Mohammad, Hendrik, Amy, Dario, Xiao and Maike, inhabitants of the Cognium and neighbors, Agnes, you've all been wonderful, inside and outside the lab. You hosted me, drove me around, translated documents, prepared coffee, shared much of the pain and the frustration, did yoga with me, and sport (and then ate burgers&cookies to compensate), went to the theater, drank a beer, and then another one, while watching the sunset at Schlachte. I wish I could have you all around, wherever I'll go next.

A thought also to Natale, because the idea of doing a PhD started on that smoky balcony at the library, where we would sit for hours, full of expectations, boredom, ambitions and youth. What followed would take up a hundred volumes and he knows how little he deserves it.

Leonardo, since the day I met you, you proved to have double the energy than a normal person has. Since that day, you've used that energy to come up with a million creative plans for fishing me out of the sofa. Thank you for listening and being there for me.
Cote and Miriam, you are two wonderful women, strong generous and funny, and this makes you two of the best people I know.
I also wish to thank all my old friends in Italy, I can't wait to hug you all again.

And Federico, for bearing with me when I'm Shantih, Ishmael or the crazy lady that lives upstairs. For being smart, confident and so relentlessly annoying. I love the way you see the world, innocent and wise at the same time, and I love what we've become – *e lo so, lo so che questo non è cipria …*

A mamma e papà, per la nostalgia dopo i viaggi in aeroporto, ogni piccola lite, le ricette per telefono e l'amore che mi dimostrate da una vita.
A Chiara, sei diventata una donna meravigliosa.

Contents

1 Introduction **2**

2 The visual system **6**

2.1 The concept of receptive field 6

2.2 Early stages of visual processing: anatomical structures and response properties 7

2.3 Primary visual cortex 8

2.3.1 Topographic organization 8

2.3.2 Response properties & the notion of sparseness 8

2.3.3 Functional organization 9

2.3.4 Intracortical connectivity 10

2.4 Further cortical processing 12

2.5 Contextual modulation 13

2.6 Spontaneous activity 14

2.7 Electrical simulation of V1 16

3 Constrained inference in sparse coding **18**

3.1 Introduction 18

3.2 Results 20

3.2.1 Extended generative model 20

3.2.2 Inference with a biologically plausible dynamics 23

3.2.3 Connection patterns and topographies 25

3.2.4 Contextual effects 28

3.3 Discussion 33

3.3.1 Relations to standard sparse coding 35

3.3.2 Connection structures 35

3.3.3 Learning rules 36

3.3.4 Neural dynamics 36

3.3.5 Contextual effects 38

3.3.6 Outlook 39

3.4 Methods 40

3.4.1 Learning and analysis of Φ and C 40

3.4.2 Simulation of the neural model 40

3.4.3 Selection of orientation contrast tuning classes 41

3.4.4 Constants and parameters 41

3.4.5 Acknowledgement 42

4 A model of spontaneous activity **43**
4.1 Introduction 43
4.2 Results 44
4.2.1 Neuronal populations: dynamics and interactions 44
4.2.2 Linear stability analysis 45
4.2.3 Network state in the marginal phase 49
4.2.4 Plausible parameter range 49
4.2.5 Dynamics of spontaneous states 52
4.2.6 Properties of spontaneous states 53
4.3 Discussion 56
4.3.1 Establishing the presence of spontaneous patterns 56
4.3.2 Implementation of a stochastic model 58
4.3.3 Possible mechanisms underlying emergence and decay of states 59
4.3.4 Matching model results to experiments 59
4.4 Methods 59
4.4.1 Detecting a spontaneous orientation-tuned state 60
4.4.2 Simulations of the stochastic models 61
4.4.3 Transition probabilities 61
5 Evoking oriented percepts **62**
5.1 Introduction 62
5.2 Results 63
5.2.1 Model description 63
5.2.2 Combining ongoing dynamics with electrical pulses 64
5.2.3 Feasibility study 66
5.2.4 Towards more complex percepts 68
5.3 Discussion 70
5.4 Methods 70
5.4.1 Numerical simulations and parameters 70
5.4.2 Detecting an orientation-tuned state 71
5.4.3 Reconstructing the evoked percepts 71
6 Conclusion **74**
6.1 Summary 74
6.2 Extensions 75
6.2.1 Full-size stimuli 75
6.2.2 Higher stages of visual processing 75
6.2.3 Spontaneous activity and criticality 76
6.2.4 Incorporating spontaneous activity into the sparse coding framework 76
6.3 Perspectives 77
Appendices **78**
A Extended sparse coding model **78**

A.1 Generalization of the model: bigger visual field 78
A.2 Contextual modulation with a bigger surround 79

B Linear stability **84**
B.1 Eigenvalues spectrum 84
B.2 Boundary between linear and marginal phase 87

Bibliography **88**

Abstract

The aim of this work is to set the basis for the development of a theoretical framework to investigate how artificial signals can be successfully introduced into primary visual cortex through electrical stimulation. This goal is approached by focusing on two different aspects of visual information processing: the contextual modulations that occur when localized visual stimuli are placed in conjunction with surround stimuli and the spontaneous activity that emerges in the absence of sensory stimulation.

Generalizing the well known standard sparse coding framework, we propose a generative model to encode spatially extended visual scenes. We show that pairing an anatomically inspired constraint (which imposes that neurons have direct access only to small portions of the visual field) to a computational coding principle (whose goal is to maximize accuracy and sparseness of stimuli-representation) is sufficient to account for a number of heterogeneous features. In particular, when trained with natural images, the model predicts a connectivity structure linking neurons with similar orientation preferences matching the typical patterns found for long-ranging horizontal axons and feedback projections in visual cortex. When subjected to contextual stimuli typically used in empirical studies, it replicates several hallmark effects of surround modulation, some of which previously unexplained, and provides a uniform explanation to contextual processing.

The dynamics of ongoing activity in primary visual cortex was investigated in a structurally simple model, where the network connectivity was chosen to mimic what we obtained from the optimization process in the sparse coding model. We used both analytical and numerical methods to study the patterns of activity that the model exhibited, identifying conditions under which biophysically realistic orientation-tuned states emerged. We quantified several properties important for comparing the model to experimental data, such as the emergence and decay probability, average persistence, localization and coexistence of different states.

In both studies, we show to what extent the properties of long-range connections between visual cortical neurons are responsible for the observed empirical facts, proposing a well-defined functional role for horizontal axons and feedback projections for contextual processing phenomena and for the generation of spontaneous tuned states.

In the last part of this thesis, we tackle more concretely the problem of inducing artificial perceptions via electrical stimulation of primary visual cortex. We present a new stimulation-paradigm which consists in monitoring the spontaneous orientation-tuned states and delivering a weak modulatory current when the cortex is in a desired state, to induce spikes in neurons that are currently close to their firing threshold. The proposed framework is tested in a structurally simple spiking neural network whose activity resembles spontaneous activity in V1. After calibrating the model to a physiologically realistic operating point, we conduct a feasibility study, investigating in particular the relations between stimulation amplitude, temporal resolution and specificity of the percept. We then show how this strategy has the potential to result in the artificial perception of an image composed by a combination of oriented features, an improvement with respect to the round phosphenes typically observed in experiments.

1 Introduction

In a normally functioning human brain, light reflecting off the objects present in the field of view enters the eyes and hits a highly structured and sophisticated organ, called the retina, where specialized receptors detect color and brightness and convert it into electric impulses. This electric signal, which contains information about the visual world, travels along the optic nerve, to the thalamus and then to the back of the head, to the primary visual cortex (V1). Over a million nerve fibers project to V1, each fanning out at least a hundred times, producing an intricate, highly recurrent network composed of a massive number of connections. Here the brain starts forming a representation (a 'code') of basic aspects of the visual signal, such as where things are in space or what shape and color they are. The information is further transmitted to higher areas where other, increasingly complex, aspects are processed – whether the objects are moving, how far away they are, whether their identity is known and which meaning they have attached – finally creating the perception that we call vision.

When the peripheral organs that carry this information to the cortex are damaged as a result of a disease or trauma, the sense of vision is lost. These injuries disrupt the process of vision at the very beginning but leave the neural machinery intact. Finding a solution to this type of blindness not only has important clinical implications, but also presents a huge and exciting theoretical challenge: Can we tap into the visual pathway, bypass the tract that isn't working and insert, through electrical stimulation, an artificial code that the brain will still interpret as a visual perception?

Even leaving aside the technological difficulties that have to be solved to reach this ambitious goal, the theoretical challenges that it poses deserve to be considered with special attention. This thesis revolves around the following two research problems:

1. Introducing electric pulses though a cortical prosthetic device in a way that mimics a true visual signal requires a deep understanding of how the brain encodes natural inputs. A spatio-temporal stimulation paradigm to elicit a given percept can only be established if we know how stable, coherent percepts of objects are formed in the cortex.
2. Cortical circuits are spontaneously active even in absence of any sensory stimulus. Since any artificial signal one plans to insert in the cortex will have to be carefully integrated with the dynamical processes that are already taking place in the brain, knowledge of the mechanisms that regulate spontaneous activity in V1 is also fundamental.

▪ ■ ▪

To tackle the first of these challenges, we investigated what pieces of information are conveyed by the activity of individual neurons and how neural populations jointly represent images. Single neurons in primary visual cortex are responsive only to a tiny region of the visual scene, which is called *classical receptive field* (CRF) and has been investigated in experiments for more

than 50 years. The region surrounding the CRF typically fails to evoke a response when stimulated alone, but can selectively modulate the neuron's response to other stimuli within the CRF (Series et al., 2003), providing a 'context'. In the real world, natural input is not confined to a small spot, but occupies the entire visual field, stimulating both the CRF of V1 neurons and their surround. Hence, in order to understand how the brain forms coherent representations of spatially extended shapes in our environment, one needs to understand how neurons integrate local with contextual information represented in neighboring cells.

Unfortunately, this integration process is anything but linear. Indeed, contextual modulations depend in a complex way on the relative contrast between stimuli in the center of the CRF and in the surround, on their orientations, directions of motion and spatial frequency. For example, the same high contrast surround stimulus placed colinearly with respect to a stimulus in the CRF of a cell, can either suppress or facilitate the activity of that cell, depending on whether the center stimulus has a high or low contrast (Polat et al., 1998).

Moreover, the objects that form a visual scene stretch out in space, bearing a certain amount of statistical regularity. In particular, the oriented edges that form their contours and outline their shapes present spatial correlations, implying that the presence of an edge in a particular location of the visual scene is informative of the presence of a second edge at different relative positions and orientations (Geisler et al., 2001). To understand how V1 neurons jointly represent complex objects, it is crucial to include these regularities into a theoretical model. But which framework is the most appropriate to investigate encoding of natural scenes?

Under natural viewing conditions the non-linear interactions between the CRF and the surround of V1 neurons produce a sparse neural activity that is energy efficient and minimizes redundancy (Vinje and Gallant, 2000). Indeed, long before these observations were made, efficiency and sparseness have been postulated as guiding principles of a neural code (Barlow, 1961). One of the most famous implementations of this idea is *sparse coding*, proposed by Olshausen and Field (1996). In their influential work, they proposed a coding strategy that maximizes sparseness or, in more concrete terms, a neural scheme in which the cortex builds a sparse representation of visual input using as few neurons as possible. As the very first application of their theory, they showed that the spatial characteristics of V1 simple cells' receptive fields emerged as fundamental components (or 'causes') of natural images. More recently, sparse coding was used by Zhu and Rozell (2013) to reproduce a variety of key effects of surround modulations. In their framework, a small localized stimulus is best explained by activating the unit whose input field best matches the stimulus. If the stimulus grows larger, other units also become activated and compete to represent it, thus inducing surround modulations. The necessary interactions between neural units are mediated by connections whose strength is inversely-proportional to the overlaps of the units' input fields. However, most of the effects observed in experiments are caused by stimuli extending far beyond the range of the recorded neuron's input fields. Hence, the mechanism put forward by this model can only be a valid explanation for a small part of these effects, covering situations in which the surround is small and in close proximity to the CRF.

In Chapter 3, we introduce a novel framework to show how sparse coding models have to be extended to better capture the cortical dynamics and the anatomical structures necessary to explain contextual processing. The novelty of the framework proposed here resides in (i) defining a way of encoding a spatially extended visual scene (i.e. more than a single small patch) that exploits statistical relations between distant features in natural images; and (ii) imposing the biophysically realistic constraint that a neural population only receives direct input from a localized region of the outside world. One of the key features of the model is the emergence

of two types of coupling structures: one that acts locally, connecting neurons that share the same input field. Another that spans a longer range, allowing for direct interactions between neurons with non-overlapping input fields – something that was missing from the standard sparse coding approach. Despite being learned in an unsupervised way from natural images, both sets of connections are consistent with anatomical findings. In particular, the long-range connections in the model link preferentially neurons with similar tuning preferences, a property that is also shared by the long-range horizontal axons that stretch for several millimeters along specific layers of V1 (Gilbert and Wiesel, 1989), and by the (even more) spatially extensive feedback connections that V1 receives from higher cortical areas (Shmuel et al., 2005).

We propose that long-range connections play a key role in the way the brain exploits statistical relations present in natural scenes to integrate local perceptions: thanks to the constraint that the input fields must have a limited size, we obtain an encoding scheme in which neurons with well-separated RF contribute to form a joint representation of a spatially extended stimulus, where collaborations between neighboring neurons are enforced by long-range connections.

▪ ■ ▪

Regarding the second of the research problems dealt with in this thesis, we investigated the mechanisms that lead to spontaneous emergence and decay of spontaneous activity patterns in visual cortex. Even if, for many practical aspects, it can be considered as a noisy background, ongoing activity in V1 actually reveals a rich spatio-temporal structure with sets of neurons that occasionally begin to fire together, either at the same time or in predictable waves (Smith et al., 2018). Those coherent activation patterns, which emerge spontaneously, involve distant functional domains with similar tuning properties over large cortical areas. As such, they appear to be very similar to the ones evoked by natural stimulation with salient stimuli, such as moving gratings with a fixed orientation, as if the cortex were to be spontaneously hallucinating a physical stimulus with a particular orientation. Typically, a spontaneous oriented state persists for anywhere from a few hundred milliseconds to a few seconds, and then the activity pattern shifts to another configuration.

The brain is thus never silent: spatio-temporally structured patterns of spiking activity constantly carry information that supports cognitive processing depending on situational demands, even in the absence of direct sensory stimulation. Many suggestions have been proposed about what could be the functional role of spontaneous activity – from memory consolidation, to development and maintenance of synaptic circuits. An intriguing idea is that spontaneous oriented states reflect expectations about possible sensory inputs (Ringach, 2009). The fact that, in the absence of sensory stimulation, the cortex *dreams* about oriented edges – the fundamental causes that make up our visual world – indicates that spontaneous activity might play an active role in vision, helping efficient processing of sensory stimuli. With the work done in this thesis, we suggest how taking advantage of spontaneous states might be helpful also in processing perceptions caused by electrical stimulation. To present this idea in a computational model, we will first focus on the following question: What mechanisms could be responsible for the generation of such coherent activity patterns?

A promising answer to this question, once again, revolves around intrinsic connectivity in visual cortex and the relation between anatomical structures and functions. The structure of spontaneous activity indeed reflects the organization of V1 lateral axons and the rules by which neurons at different cortical locations and with different preferences for stimulus attributes connect to each other. To reproduce the experimentally observed cortical states, previous

modelling work (Goldberg et al., 2004; Blumenfeld et al., 2006) proposed a simplified V1 neural network that, once tuned to a particular parameter regime, could wander among a ring of attractors, each representing a particular orientation map, where transitions across similar angles are more likely to occur. Many questions, however, still need an answer, especially in light of the most recent experimental findings. For example, the dynamics observed in experiments is not only characterized by smooth transitions in orientation space, but also by abrupt jumps. Another aspect that was not thoroughly explored is the presence of *mixed* states, that is states that are composed of different orientation maps in different cortical regions, found occasionally in recordings under anesthesia (O'hashi et al., 2017), or localized states, whose lateral spread spans only a few hypercolumns, matching the description of activity patterns observed in awake animals (Omer et al., 2018; Smith et al., 2018), where widespread activity is more rare. To explain the emergence and decay of spontaneous states, a more comprehensive modeling approach is necessary. We address these issues in Chapter 4, where we present a model whose architecture is inspired by the neural network derived from the implementation of sparse encoding of spatially extended visual scenes (Chapter 3). The properties of activity maps generated by the model are analyzed both analytically and numerically and are compared to experimental data.

▪ ■ ▪

The goal of building a visual prosthesis dates back to the work of Brindley and Lewin (1968) and Dobelle and Mladejovsky (1974), scientists that studied *phosphenes*, the perceptual sensations evoked by electrical stimulation of the occipital cortex. Almost all verbal reports of phosphenes reported by both sighted and blind patients who received stimulation agree in describing them as bright and round. This appearance most likely results from the broad activation of large population of neurons: given the fixed geometry of electrode arrays, microstimulation targets cells with all possible orientation preferences, resulting in an unspecific percept. Is it possible to improve the specificity of the artificial percept evoking, for example, an elongated feature?

More concretely, by constantly monitoring which information is dynamically represented and processed in visual cortex, can we stimulate a neural network when it is already in a desired state and evoke a percept with the corresponding orientation? Suppose we can observe different cortical locations being tuned to different orientations, can we take advantage of subsequent activity-configurations to evoke a combination of oriented features resembling a more complex object?

These hypotheses are tested in Chapter 5 of this thesis. There, we use the knowledge gained from the mean-field model of spontaneous activity to implement a second, more plausible spiking network that is able to generate realistic spatio-temporal patterns of activity. In this framework, we motivate the idea behind the proposed stimulation paradigm with numerical investigations and perform a feasibility study, investigating in particular the relations between stimulation amplitude and temporal resolution, and specificity of the percept.

2 The visual system

In this chapter we will present a brief introduction to the visual system, with a particular emphasis on primary visual cortex, its structure and its computational properties. The biological and theoretical background that will be presented is not intended to be exhaustive, but is instead limited to what is relevant to understand the content of this thesis.

2.1 The concept of receptive field

A first step for understanding how sensory information is processed by the visual system is to consider the responses of individual neurons and, in particular, to determine which stimuli more effectively drive them.

Neurons in the visual system respond to light stimuli in restricted areas of the visual field. Within such areas, there are regions where illumination brighter than the background light-intensity enhances firing, and other regions where darker illumination enhances firing. The position of those regions, together with their spatial arrangement, determines the selectivity of the neuron to different inputs and it is called the *receptive field* (RF). A simple, concrete example of the RF of neurons in early areas of the visual system is presented in Fig. 2.1, but we will give more thorough descriptions in the following sections.

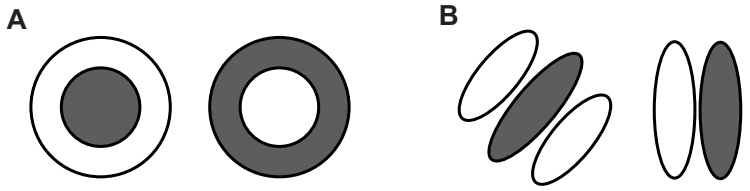


Fig. 2.1: Receptive field diagrams for retinal ganglion cells (**A**) and V1 neurons (**B**). Dark regions are termed OFF regions and light circles ON regions.

The concept of RF is central to sensory neurobiology, since it provides knowledge of where and how one has to stimulate a cell to make that cell respond. Its precise characteristics depend on how it is measured. The classic method to determine the position and extent of the RF of a neuron in a visual area is to present discrete stimuli at different locations on the retina: The region that yields deviations in firing rate above the background activity level is referred to as the 'classical receptive field' (CRF).

A consequence of this definition is that stimuli presented outside the CRF are unable to elicit a response. However, as described in detail in Section 2.5, probing the surrounding regions of V1 cells' CRF can significantly affect responses to stimuli presented inside their RF. Formally, the concept of a receptive field is captured in a model by including a linear filter as its first stage, where filtering involves multiplying the intensities at each local region of an image (the value of each pixel) by the values of a filter and summing the weighted image intensities. This

linear operation, though, fails to predict the response of cortical neurons to arbitrary stimuli, in particular to large visual scenes that extend beyond the limit of their CRF. Thus, the notion of classical receptive field alone is not sufficient to explain how the brain integrates information from other cells, neighboring or distant, to form coherent representations of the visual world.

2.2 Early stages of visual processing: anatomical structures and response properties

The first stage of the visual system is the eye. The light reflected from objects in the world, after passing through the cornea and the lenses, falls into the retina (Fig. 2.2). The retina contains an array of specialized receptors reacting to electromagnetic waves in the visible spectrum. These receptors, through a process called phototransduction (Callaway, 2005), transform physical properties of light into electrochemical signals. Such signals are initially processed by a network of nerve cells and then passed on to the retinal ganglion cells.

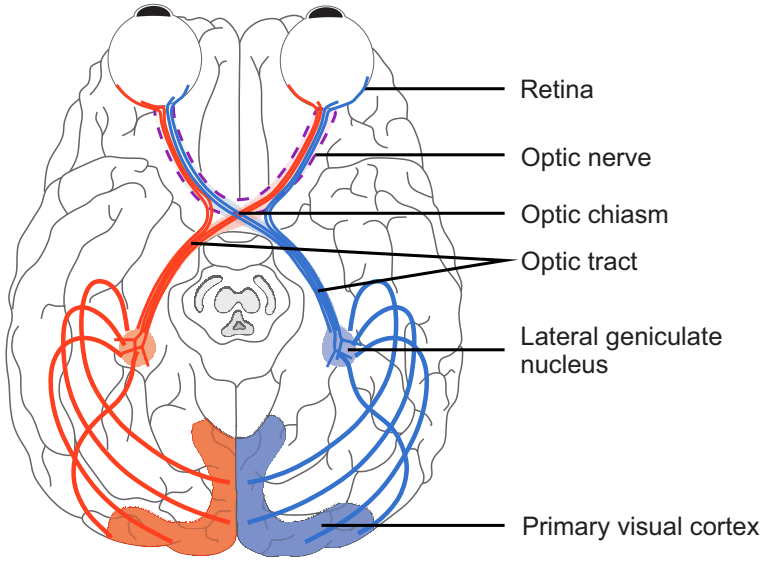


Fig. 2.2: Pathway from the retina through the lateral geniculate nucleus (LGN) of the thalamus to the primary visual cortex in the human brain.

Here the light signal undergoes a first important transformation, resolution-downsampling: This term refers to the fact that, on average, each retinal ganglion cell receives inputs from about 100 rods and cones, thus performing a dramatic compression of the information delivered to the brain. However, these numbers vary greatly as a function of retinal location: in the fovea (center of the retina), a single ganglion cell communicates with as few as five photoreceptors; in the periphery, a single ganglion cell receives information from many thousands of photoreceptors. Thus compression occurs with minimal loss of information so that detailed visual information is preserved.

Retinal ganglion cells have concentrically organized spatial RF's with either ON or OFF centers, as schematically shown in Fig. 2.1 (A). An ON region is defined as a region in which a bright light evokes a positive response and a dark light evokes a negative response. This means that light increments falling in the center of an ON cell's RF increase firing, while light increments in the surround reduce firing, and *vice versa* for OFF center cells (Hartline, 1938; Kuffler, 1953; Hubel and Wiesel, 1962). This center-surround structure allows ganglion cells to respond to

local variations of light intensity – signaling the illumination of a center location relative to that of the surround – and it is the basis of edge detection. Thus, the operations that the retina performs on the images consist in enhancing the edges of objects within its visual field.

The reasons for such organization may reside in metabolic efficiency and redundancy reduction. With respect to the pixel-based representation of visual inputs operated by the photoreceptors, the representation of ganglion cells in terms of boundaries of objects avoids the cost (in terms of spikes) of signaling uniform regions and makes it possible to perceive a large object only through those cells that are confined to the borders.

The axons of retinal ganglion cells, bundled in the optic nerve, leave the eye to project to the lateral geniculate nucleus (LGN) (Callaway, 2005), a structure of the thalamus organized in 6 layers (Fig. 2.2). Most cells in the LGN exhibit the same ON/OFF center behavior as that retinal ganglion cells that provide input to them. The LGN relays the incoming information towards the striate cortex. The functional role of the LGN (and of the thalamus generally) is still unclear, though there is evidence that it is responsible for temporal decorrelation at different spatial and temporal scales (Dong and Atick, 1995) and for attentional modulation (O'Connor et al., 2002).

2.3 Primary visual cortex

Located in the occipital lobe of the cerebral cortex (Fig. 2.2), the primary visual area (V1) is the first stage of cortical processing of visual information. It receives its main visual input from the LGN, sends its main output to subsequent cortical visual areas, and is traditionally divided in 6 horizontal layers.

In many species, area V1 lies at least partly on the cortical surface, and is therefore accessible for various imaging methods. For this reason, over the past 50 years it has been extensively investigated. It is now one of the best understood areas of the cerebral cortex and constitutes a prime workbench for the study of cortical circuits and of computations: we understand the nature of its main inputs, we know what stimuli make its neurons fire and we can easily control many properties of those stimuli.

2.3.1 Topographic organization

A striking feature of the visual system is that the visual world is mapped onto the cortical surface in a topographic manner: neighboring points in a visual image evoke activity in neighboring regions of visual cortex. Topographic representation of the visual world occurs in the visual system at many levels, starting with the image that forms in the retina and it is known as *retinotopy* – a notion that resembles continuity in mathematics. Retinotopic mapping is maintained also in the LGN, in V1 and in many other visual areas (Van Essen et al., 1984) and it allows mapping RF positions in the retina to the corresponding RF positions on the cortical surface.

2.3.2 Response properties & the notion of sparseness

In addition to stimulus position, V1 neurons are selective for a number of attributes, including orientation, direction of motion, spatial and temporal frequency and, in many species, for binocular depth and color (Hubel and Wiesel, 1962; De Valois and De Valois, 1980). Given this elaborate selectivity, V1 cells display a richer variety of receptive field shapes than cells in upstream visual areas (Ringach, 2002).

Most RFs in V1 are characterized by ON and OFF subregions and have an elongated shape, as schematically exemplified in Fig. 2.1 (B). However, the exact number and exact position of subregions or the precise aspect ratio can vary considerably.

A mathematical approximation of the spatial profile of the receptive field of V1 cells is provided by Gabor functions, sinusoidal plane waves with a Gaussian envelope, described by the equation

$$g(\theta, \lambda, \sigma_x, \sigma_y, x_0, y_0, \psi) \propto \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right) \cos\left(2\pi\frac{y}{\lambda} + \psi\right)$$

$$x = +x_0 \cos(\theta) + y_0 \sin(\theta)$$

$$y = -x_0 \sin(\theta) + y_0 \cos(\theta),$$

In this expression, the orientation of the sinusoidal carrier θ represents the preferred orientation of the neuron, the wavelength λ represents the inverse of its preferred spatial frequency and the phase ψ its preferred phase, while the standard deviations σ_x and σ_y determine the size and the aspect ratio of the RF.

While some cells are activated only by a specific phase of such a grating (originally termed 'simple cells' (Hubel and Wiesel, 1962)), other cells (termed 'complex') respond to gratings regardless of their phase. Complex cells respond optimally to moving stimuli that move in specific directions. Thus, to obtain a complete picture, the spatial receptive field structure should be complemented with the temporal aspects of stimulus changes, leading to spatiotemporal receptive fields.

Due to the characteristics of their RFs, neurons in V1 preferentially respond to local image patches containing oriented elements, such as bars or gratings. Because of this, they are normally considered as edge detectors or, in other words, they create a *sparse code* for edges.

To clarify what this means, it might help to consider the coding of edges at various levels of the visual system. At early processing stages, information about the presence, orientation and location of an edge is carried by means of population codes: A recording from a single photoreceptor or a single ganglion cell would yield little information about the presence or position of the edge. In V1, remarkably, an edge at a particular location creates activity in only relatively few neurons (hence the term *sparse*) – the neurons tuned to the appropriate orientation at the appropriate retinal location – and recording from the right individual neurons would yield considerable information about the presence of the edge.

2.3.3 Functional organization

Neurons in the primary visual cortex are arranged vertically into columns of neurons that have similar functional properties. For example, neurons in different layers of the cortex, but with similar tangential position (e.g. whose cell bodies fall within 30 – 50 µm of a line drawn perpendicular to the pial surface) might respond primarily to stimuli that have a certain orientation (e.g. within approx. 10 degrees) and are perceived by the same eye (Purves et al., 2001).

In addition to following a retinotopic organization, V1 neurons are arranged in so-called visual maps according to their tuning properties, so that stimulus attributes are mapped in an orderly fashion across the brain. For example, when the distribution of orientation selectivity in a plane parallel to the cortical surface is inspected, preferences usually rotate either clockwise or counter-clockwise at a roughly constant rate; the direction of rotation will typically continue unchanged for 1 – 2 mm and then reverse unpredictably. The result is an orientation map

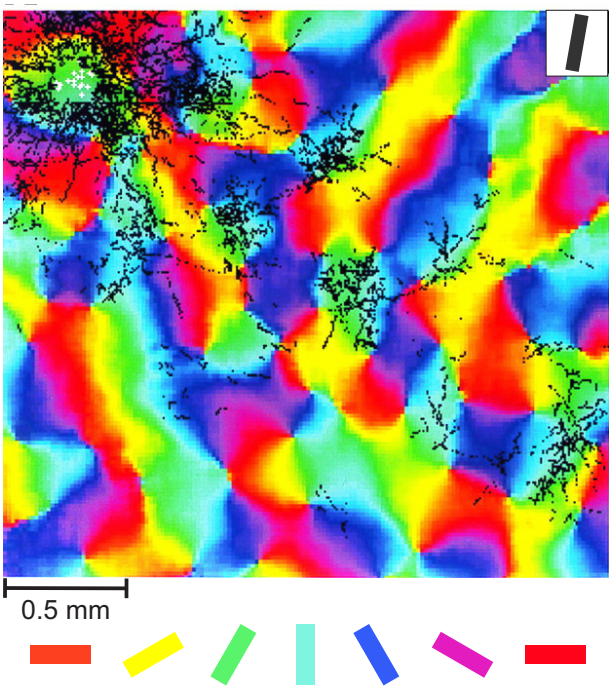


Fig. 2.3: Orientation maps and horizontal connections. Each point on the cortical surface is color coded according to the orientation preference measured at that location, as indicated by the oriented bars. The black symbols show the terminations of the long-range horizontal axons (*boutons*) of a pre-synaptic cell whose location is indicated in white and whose orientation preference is depicted in the top-right corner. The distribution of boutons indicate that, at long distance from the cell body, connections are made preferentially between sites with similar orientation preferences. Modified from (Bosking et al., 1997).

of periodically arranged columns within which, at every 1 – 1.5 mm, the same orientation is encountered (Fig. 2.3). Orientation maps have been found in primary visual cortex of primates and carnivores such as macaque monkeys (Blasdel and Salama, 1986; Blasdel, 1992b), tree-shrews (Bosking et al., 1997), ferrets (Rao et al., 1997) and cats (Hubel and Wiesel, 1962; Löwel et al., 1988; Bonhoeffer and Grinvald, 1991; Ohki et al., 2005, 2006). The cortical region that encompasses a complete cycle of orientations is called *hypercolumn*, a term originally coined by Hubel and Wiesel (Hubel and Wiesel, 1974) to denote a functional unit processing all the information coming from a specific location in the visual field.

An analogous functional architecture, similarly striking in organization and precision, has been identified for ocular dominance (Hubel and Wiesel, 1968; Blasdel, 1992a) (typically composed of alternating stripes where neuronal responses are dominated by one input or the other), spatial frequency (Issa et al., 2000; Nauhaus et al., 2012), direction (Weliky et al., 1996; Ohki et al., 2005), color (Landisman and Ts'o, 2002) and disparity (Kara and Boyd, 2009) but it will not be discussed further, since it is not directly connected with the core of this thesis.

2.3.4 Intracortical connectivity

Area V1 is divided in 6 horizontal layers, with a characteristic distribution of inputs and outputs across layers (Douglas and Martin, 1998). Feed-forward inputs from LGN arrive in layer 4, the

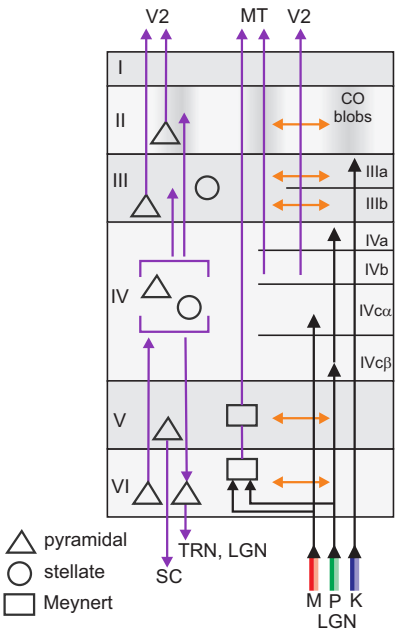


Fig. 2.4: Connectivity circuits. Sketch of major connections, afferents, efferents, and cell types in primary visual cortex. Inputs from LGN are shown as black arrows. Dark gray symbols depict some of the major cell populations found in the different layers, together with selected ascending and descending fibers shown as violet arrows. Horizontal fibers are indicated by the orange arrows. In layer II, CO blobs are indicated by darker shading. Redrawn from (Kretzberg, 2013).

'input layer', with collaterals to layer 6 while feedback inputs from other cortical areas arrive mostly in superficial and deep layers. Feed-forward outputs to other cortical areas depart from layers 2/3 while feedback outputs to the thalamus to other subcortical targets depart from layers 6 and 5 respectively (Angelucci et al., 2002).

Within V1 itself, two basic types of intracortical connections can be identified based on their distribution relative to the cortical surface (reviewed in (Fitzpatrick, 1996)). The most dense, and the first one to be identified with anatomical techniques, includes axons that travel perpendicular to the pial surface and provide much of the *vertical* communication between layers (Valverde, 1971; Lund, 1973). This local circuit (shown in Fig. 2.4) operates at sub-millimeter dimensions, with axons' terminal fields that arborize with relatively little lateral spread (roughly 0.5 mm).

The second type, in order to be identified and characterized, required the development of more sensitive anatomical tracing techniques (Gilbert and Wiesel, 1979). It consists of a system of *horizontal* axon arbors extending over long distances (2-8 mm) parallel to the pial surface (Rockland and Lund, 1983; Gilbert and Wiesel, 1983, 1989). This circuit operates over a longer range and is mediated by the horizontally spreading axons of excitatory pyramidal neurons.

Short-range connections form a massive network of axonal and dendritic arbors – the most prominent vertical pathways go from layer 4 to layers 1-3, from layer 6 to layer 4 and from layers 2/3 to layer 5 (Amir et al., 1993). Their strength is a decreasing function of lateral separation, largely radially symmetric (Das and Gilbert, 1999). There is a general consensus that local connections are relatively independent of orientation preference, although recent findings in mouse visual cortex reveal that connectivity is structured also on a local scale: Neurons with the same preference for oriented stimuli connect at a higher rate than neurons with orthogonal

orientation preference (Ko et al., 2011) and even higher if their RFs are aligned along the axis of their preferred orientation (Iacaruso et al., 2017).

Long-range horizontal connections are primarily made via cells in layers 2/3, 5 and 6 (Gilbert and Wiesel, 1983; Casagrande and Kaas, 1994; Rockland and Lund, 1983). These connections contact predominantly excitatory (≈ 80%) but also inhibitory (≈ 20%) neurons, they have the tendency to arborize in preferred sites, forming distinct axonal clusters of 200 – 300 μm in diameter, and link preferentially cortical domains of similar functional properties, such as orientation preference (Ts'o et al., 1986; Gilbert and Wiesel, 1989; Weliky et al., 1995; Bosking et al., 1997; Malach et al., 1993), ocular dominance columns (Malach et al., 1993) and CO compartments (Yoshioka et al., 1996). In the tree shrew (Bosking et al., 1997), cat (Schmidt et al., 1997) and new world primates (Sincich and Blasdel, 2001), lateral connection in layers 2/3 are anisotropic, and their axis of anisotropy has been shown to be collinear in space with the orientation preference of the neurons of origin. In contrast to feed-forward thalamic axons, horizontal axons do not drive their target neurons, but only elicit sub-threshold responses (Hirsch and Gilbert, 1991; Yoshimura et al., 2000), thus having a modulatory influence.

2.4 Further cortical processing

In the cortex, besides V1, a number of anatomically distinct areas contain neurons that respond selectively to visual stimulation. These areas, located in the temporal and parietal lobes, are mutually interconnected and form a complicated network where information flows along feed-forward and feedback connections (Felleman and Van, 1991), even though theoretical descriptions almost always present them as a 'hierarchy'.

One popular conceptualization of how the primary visual cortex is functionally linked to the extra-striate areas is that there are two main pathways (Fig. 2.5) by which information travels from V1 to the surrounding visual areas, a ventral and a dorsal pathway (Mishkin et al., 1983). The former runs through the temporal lobe, in visual areas 2 (V2), visual area 4 (V4) and inferior temporal cortex (IT), and is associated with object recognition and form representation. The latter runs through the parietal lobe, in areas V2, V3, and middle temporal area (MT), and is involved with perceiving motion and spatial relationships between objects in the visual field.

A sequential routing and processing along both streams is, of course, a simplification, since the two streams are not parallel and cross-talk between the two exists. Nevertheless, both streams exhibit hierarchical characteristics. As one proceeds downstream to higher visual areas, the response latencies, as well as the complexity of stimulus selectivity increases. Receptive fields tend to become more and more complex and their description increasingly elaborate. For example, some IT neurons are size or location invariant and thus respond similarly to objects at different distances or irrespective of where they are in the visual field, and some others respond to complex stimuli of specific shapes combined with specific color and texture, such as faces (Desimone et al., 1984).

The numerous operations that the brain has to perform on a given visual scene to pool and bind together local features into global coherent percepts in order to extract task-relevant information, assigning a behavioral meaning to objects or generally make sense of it, are currently a subject of intense research. In the following Section, we consider how, at the level of primary visual cortex, the problem of integrating local information from distant cells is highly non-trivial.

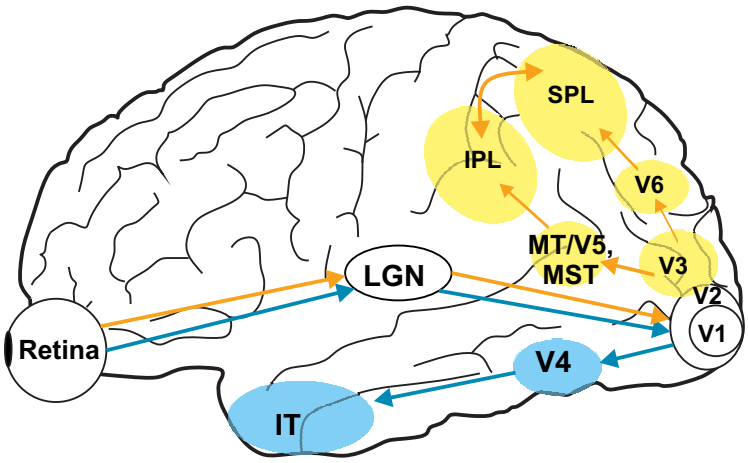


Fig. 2.5: Ventral and dorsal pathways. Separation of visual information processing into two major streams in higher cortical areas. Areas within the same pathway are represented with the same color, with blue shades for the ventral stream and yellow shades for the dorsal stream.

2.5 Contextual modulation

As already explained in Section 2.1, neurons in V1 respond to presentation of visual stimuli within a localized region of space, the neuron's receptive field. Presentation of similar stimuli outside of this region typically does not evoke a response from the neuron but can modulate (suppress or facilitate) the neuron's response to stimulation of its CRF.

Since Hubel and Wiesel (Hubel and Wiesel, 1968) first discovered that the firing rate of some cells decreased despite being presented with increasingly larger stimuli, it was clear that the concept of CRF was not sufficient to describe completely the behavior of neurons in V1.

Electropyhsiological studies performed in the past 40 years have revealed a multitude of phenomena which have been termed *non-classical* RF effects (NCRF). Typically these experiments are conducted with two stimuli, a center stimulus and a surround stimulus. The center stimulus is placed in a location of the visual field in retinotopical correspondence to the CRF of the probed neuron and matches one or more of its tuning properties. The surround stimulus is placed in the surrounding of the center stimulus, either in a concentric configuration, or placed sidewise. Changes in the neuron's response are observed upon varying systematically one or more attributes (e.g. orientation, size, contrast) of either one of the stimuli in different conditions, such as center alone, center and surround or surround alone.

The observed phenomena in general depend in a complicated manner on various parameters like the stimulus configuration, contrast, and geometry, revealing how a far more complex processing than plain linear filtering is carried out in primary visual cortex.

Specifically, the effects of surround stimulation are selective for orientation and direction. Maximal modulations are generally observed when center and surround stimuli have the same orientation (Levitt and Lund, 1997; Sengpiel et al., 1997; Sillito et al., 1995; Walker et al., 1999; Kapadia et al., 1995, 2000; Knierim and Van Essen, 1992; Chen et al., 2001; Nelson and Frost, 1985; Polat et al., 1998) and similar maximal effects are found for stimuli of similar spatial frequencies (DeAngelis et al., 1994; Chao-Yi and Wu, 1994; Walker et al., 1999) and speed (Chao-Yi and Wu, 1994). Regarding the sign of the effects, modulations are mostly inhibitory

(Jones, 1970; Sengpiel et al., 1997; Walker et al., 2000). However, excitatory effects are also known: in most cases they appear for discrete stimuli (bars, Gabor patches) presented at the end zones of the cRF or when the center and surround are coaxially aligned and well separated. The contrast of the center stimulus relative to the cell's contrast threshold appears to control the sign of the modulation (Levitt and Lund, 1997; Mizobe et al., 2001; Polat et al., 1998; Sengpiel et al., 1997; Toth et al., 1996); sometimes the same surround stimulus can facilitate the response to a low-contrast center stimulus and suppress the response to a high-contrast center stimulus (Polat et al., 1998; Chen et al., 2001). Finally, the strength of the modulation decreases with spatial separation between center and surround patches, but can still be observed for distances up to 12 degrees of visual angle (Mizobe et al., 2001).

One long-standing and highly influential idea is that the ncRF provides context for stimuli appearing in the cRF, enhancing the ability of neurons to detect or discriminate orientation and motion discontinuities, textures and contour curvature or even facilitate target selection by pop-out mechanisms.

Many of the investigated phenomena require integration of visual signals arising from regions in visual space that are well segregated and therefore must depend on interactions between neurons whose cRF are non-overlapping. Such contextual modulations could be conveyed through intra-cortical lateral connections or feedback from higher cortical areas or a mixture of both. In Chapter 3 we consider a model that is able to reproduce several well established ncRF effects and we will discuss possible neural circuits that could be responsible for generating them.

2.6 Spontaneous activity

Even without any active afferent stimuli coming to the sensory areas, the brain is known to display spontaneous activity whose nature and origin is still a matter of debate. This spontaneous activity, also called ongoing activity, is by definition the running activity of the brain when no particular stimulus is being processed or when no particular (or at least no known) actions are performed. In the absence of a visual stimulus, we can assume that any activation pattern displayed by the visual cortex is determined by intrinsic properties of cortical networks. Therefore, investigating those patterns would allow us to have a view on cortical information processing which is potentially unobstructed by an imposed external input.

Spontaneous activity has been shown to contribute to trial-to-trial variability of subsequent evoked sensory responses (Arieli et al., 1996) and has been related to processing and replay of sensory experience (Karlsson and Frank, 2009; Wilson, 2010), reorganization of synaptic weights (Wang et al., 2011), memorization of sensory events (Deuker et al., 2013; Abel et al., 2013) and understanding the mechanisms that generate it is likely to help in understanding the fundamental principles behind cortical processing.

The nature of this activity depends on the behavioural state, such as wakefulness/sleep, level of alertness, expectations, or even on the sequence of the preceding stimuli. In awake animals, however, ongoing background activity possesses certain features that are common in many species and many cortical areas: it appears irregular and rather sparse (Hubel, 1959; Lin et al., 2006; Ferezou et al., 2006; Greenberg et al., 2008), with neurons firing spontaneously at relatively low rates.

Multielectrode recordings from various regions of the visual system, including the retina (Meister et al., 1991), LGN (Weliky and Katz, 1999), V1 (Chiu and Weliky, 2001) and extra-striate areas (Destexhe et al., 1999) have shown that, despite the apparent irregularity in firing

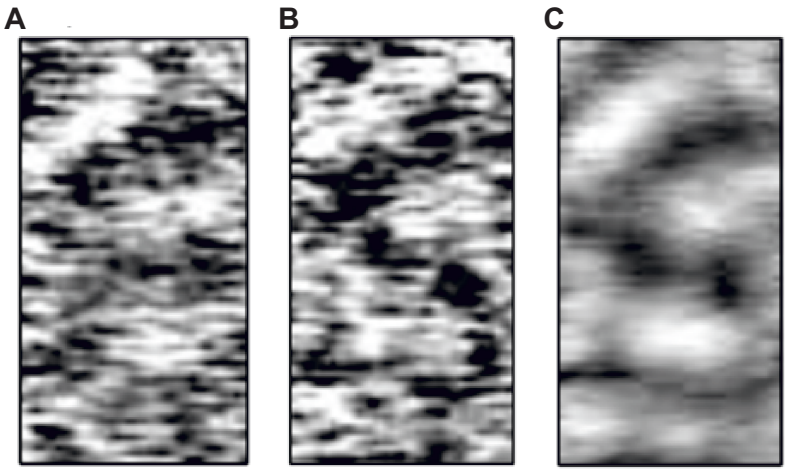


Fig. 2.6: (**A**) Instantaneous pattern of spontaneous activity. (**B**) Instantaneous pattern of activity evoked using a full-field grating stimulus with a vertical orientation. (**C**) Orientation maps obtained by averaging many instantaneous evoked patterns with the same vertical orientation. Redrawn from (Kenet et al., 2003).

observed at the level of single cells, spontaneous activity does not appear noisy or random, but instead shows a high degree of coherence in both spatial and temporal dimension. In particular, in primary visual cortex, bursts of activity tend to occur simultaneously over large cortical distances (several millimeters), with a typical duration of tens to few hundreds of milliseconds (Chiu and Weliky, 2001).

Study of ongoing activity advanced considerably with the refinement of VSDI (Voltage Sensitive Dye Imaging), an imaging technique that enables one to visualize the activity of neural populations in a large patch of cortex with high temporal resolution (Arieli et al., 1995; Grinvald and Hildesheim, 2004). With VSDI, one obtains temporal frames, each representing an activation pattern, or *cortical state*, either in the presence or absence of a visual stimulus. In particular, averaging the optical signal over many presentations of a full-field grating with a fixed orientation yields a so-called single orientation map. Combining extracellular recordings with VSDI in anesthetized cats, Tsodyks and colleagues (Tsodyks et al., 1999) realized that the spontaneous action potentials of V1 neurons very often occur when the instantaneous cortical states resemble the single orientation map obtained from stimuli whose orientations match the tuning properties of such neurons. This suggestive relation between ongoing activity and the functional architecture of visual cortex was further explored in (Kenet et al., 2003): analysing the dynamics of cortical states, it emerged that spontaneous activity reflects a dynamic switching between a set of intrinsic states, many of which correspond closely to single orientation maps. More recently, similar experiments were conducted (O'hashi et al., 2017; Omer et al., 2018; Smith et al., 2018), in order to better characterize the properties of such oriented states. When one of such pattern emerges, it spans several hypercolumns and lasts on average 200 ms. In orientation space, the state-switching is often smooth, i.e. oriented states are often followed by a state corresponding to a proximal orientation, but sometimes abrupt changes, in which the cortex tends to switch to orthogonal orientations, are also observed (O'hashi et al., 2017).

The functional significance of such findings is not clear, since the described cortical states occur in sedated animals. Despite investigation in awake monkeys did not clearly reveal such *globally* organized activity patterns (Omer et al., 2018), spontaneous correlated activity involving

orientation-domains was still observed in monkeys (Omer et al., 2018) and ferrets (Smith et al., 2018) without anesthesia. In these cases, spontaneous activity patterns consist of a distributed set of active domains which become active either simultaneously or in a spatiotemporal sequence, spreading across the imaged area within a few hundred milliseconds.

The mechanisms responsible fo generating these spontaneous events are still not clearly understood. Even though a causal role for retinal and thalamic feedforward inputs in establishing correlated modular structures cannot be ruled out, experimental evidence suggests that such complex states are shaped and expressed through intrinsic cortical mechanisms.

Theoretical investigations have corroborated the idea that lateral orientation-specific connections have a great relevance to the large-scale spatial organization of both sensory representation in primary sensory cortices and spontaneous activity (Ernst et al., 2001; Goldberg et al., 2004; Blumenfeld et al., 2006). However, to explain the emergence and decay of spontaneous states a more comprehensive modeling approach is necessary, where the role of noise might be crucial. Some spatial and temporal aspects of spontaneous dynamics, indeed, were not thoroughly investigated; those include, for example, the presence of *mixed* states, that is states that are composed of different orientation maps in different cortical regions, or of localized states, whose lateral spread is less than the whole imaged area. Mosaic and localized states are more compatible with the activity patterns observed *in vivo* (Omer et al., 2018; Smith et al., 2018), therefore it is interesting to understand which interactions determine them, especially if we wish to exploit ongoing cortical states to insert artificial signals.

2.7 Electrical simulation of V1

Among the technologies that allow one to interfere with the natural states of the brain, electrical stimulation is one of the most powerful techniques for establishing a direct contribution of neuronal activity to different levels of visual processing, in particular to visual perception. Electrical stimulation involves the introduction of electrical current into a small cortical region either through an electrode placed on the cortical surface ('epidural/subdural cortical stimulation') or a microelectrode inserted into the cortical matter ('intracortical microstimulation') (Doty, 1965).

The ability to detect external electrical stimulation has been extensively characterized in primates' visual cortex (for recent and exhaustive reviews see (Histed et al., 2013; Tehovnik and Slocum, 2013) or (Cicmil and Krug, 2015)). In detection tasks, animals report the presence or absence of electrical stimulation within a given time period, for example, by pressing a lever or making a saccade to an appropriate target. With little prior training to recognize electrical stimulation, monkeys can reliably detect strong electrical stimulation of area V1. However, extensive training, numbering thousands of trials, is necessary to achieve stable low detection thresholds (i.e. below 10μA for microstimulation and around 0.1 – 1mA for subdural stimulation). Once learned, monkeys can generalize the detection to any region within V1.

An intriguing discovery made in the late sixties revealed that cortical surface stimulation of V1 in humans produces the sensation of a small point of light, called phosphene (Brindley and Lewin, 1968; Dobelle et al., 1976; Schmidt et al., 1996; Dobelle, 2000; Pollen, 2004; Dobelle and Mladejovsky, 1974; Bak et al., 1990; Murphey et al., 2009) and that the apparent locations of phosphenes with respect to the stimulating electrode agree with retinotopic maps of the visual field in cortex. Descriptions of perceived phosphenes were not uniform across tested patients. While some studies reported a lack of colour sensation upon stimulation (Brindley and Lewin, 1968; Lee et al., 2000), in other cases the chromatic effects of phosphenes were

vivid reds, blues or greens or 'unreal' colours. In most cases, phosphenes had a round shape, but occasionally patients have reported elongated phosphenes (Brindley and Lewin, 1968).

Being able to predict the effects of electrical stimulation is crucial for the development of cortical prosthetic devices, with which one can artificially manipulate signals in those parts of the brain that control areas of the body where the function has been lost. This would have important clinical applications, for example, in restoring sight in patients with acquired blindness where the eyes or the optic nerve are damaged. Before cortical prostheses can become a viable option, it is necessary to understand how to generate percepts that are more complex than a single spot of light.

A first possibility to achieve this, would be to stimulate regions along the visual hierarchy whose neurons are selective for more elaborate combinations of features than V1 neurons. However, it is generally more difficult to evoke detectable sensations with electrical stimulation of extra-striate visual areas using surface electrodes (Murphey et al., 2009; Lee et al., 2000). Even when detectable sensations are elicited, reports differ regarding the content of the evoked sensation. In some studies, patients reported sensations of 'complex forms', such as faces or visual scenes from memory, while in other studies only simple form sensations were evoked. These differing results may be due to individual differences in extra-striate function between patients. But, they might also reveal current limitations in our understanding and control of the effects of direct electrical stimulation on the volume of brain tissue below a cortical surface electrode.

A second possibility would rely on in the improvement to chronic implantation techniques. Intracortical microelectrodes, albeit more invasive that subdural electrodes, might provide a more effective prosthetic approach, targeting specific subregion of the cortex and using smaller currents. Using array with a high number of electrodes with a sufficiently fine spatial resolution, visual information could be conveyed through increasingly complex patterns of electrical stimulation (subjects can typically resolve phosphenes produced by electrodes separated by as little as 500μm (Bak et al., 1990; Schmidt et al., 1996)). However, combining phosphenes to obtain more detailed images is not so straightforward since concurrent stimulation of multiple sites in visual cortex produces multiple phosphenes that do not combine to form a coherent shape (Dobelle et al., 1976; Schmidt et al., 1996). One explanation for the failure of the conventional electrical stimulation paradigm is the unnatural activity that it evokes in cortex. When viewing natural scenes, only a small fraction of neurons in early cortex are active. In contrast to selective activation of neurons produced by real visual stimuli, electrical stimulation activates an effectively random set of neurons in the immediate region of the electrode (Histed et al., 2009). Even though this electrical activation can result in a roundish percept, the non-selective activation of spatially contiguous neurons might not propagate to higher areas to produce complex percepts as normally occurs with natural vision.

A third, less explored possibility would consist of exploiting the ongoing dynamics of primary visual cortex – the spontaneous activation patterns, reflecting functional connectivity and activating populations of neurons with similar orientation preferences, described in Section 2.6. Instead of targeting a large populations of neurons through intracortical microstimulation, that would activate neurons with all possible orientation preferences and thus result in a round and unspecific percept, one could wait until the emergence of a desired orientation state and use a weak modulatory current to induce spikes in neurons being spontaneously close to their firing threshold. This would result in a percept of an elongated oriented feature, perhaps easier to combine with other oriented features into complex shapes. We explore this idea in Chapter 5.

3 Constrained inference in sparse coding: contextual effects and neural dynamics

In this Chapter we will build a novel framework for contextual processing in the visual system. In particular, we will propose a generative model to encode spatially extended visual scenes, generalizing the standard sparse coding model by including spatial dependencies among different features. After deriving a physiologically realistic inference scheme and mapping it to a network where synaptic interactions match the properties found for long-ranging connections in visual cortex, we will show that our model replicates several hallmark effects of surround modulation, suggesting a well-defined functional role for horizontal axons and feedback projections.

3.1 Introduction

Single neurons in the early visual system have direct access to only a small part of a visual scene, which manifests in their 'classical' receptive field (cRF) being localized in visual space. Hence for understanding how the brain forms coherent representations of spatially extended components or more complex objects in our environment, one needs to understand how neurons integrate local with contextual information represented in neighboring cells. Such integration processes already become apparent in primary visual cortex, where spatial and temporal context strongly modulate a cell's response to a visual stimulus inside the cRF. Electrophysiological studies revealed a multitude of signatures of contextual processing, leading to an extensive literature about these phenomena which have been termed 'non-classical' receptive fields (ncRFs) (for a review, see (Angelucci and Shushruth, 2013; Series et al., 2003)). ncRF modulations have a wide spatial range, extending up to a distance of 12 degrees of visual angle (Mizobe et al., 2001) and are tuned to specific stimulus parameters such as orientation (Sengpiel et al., 1997). Modulations are mostly suppressive (Walker et al., 2000), although facilitatory effects are also reported, especially for collinear arrangements where the center-stimulus is presented at low contrast (Polat et al., 1998) and for cross-orientation configurations (Sillito et al., 1995; Levitt and Lund, 1997). However, there is also a considerable variability in the reported effects, even in experiments where similar stimulation paradigms were used: for example, (Polat et al., 1998) found iso-orientation facilitation for low center stimulus contrasts, whereas another study (Cavanaugh et al., 2002a) did not report facilitation at all, regardless of the contrast level. A further example (Sillito et al., 1995) found strong cross-orientation facilitation, while (Levitt and Lund, 1997) reports only moderate levels of cross-orientation facilitation, if at all. These discrepancies might be rooted in differences between the experimental setups, such as the particular choice of center/surround stimulus sizes, contrasts, and other parameters like the spatial frequency of the gratings, but might also be indicative of different neurons being specialized for different aspects of information integration.

From the observed zoo of different effects in conjunction with their apparent variability, the question arises if explanations based on a unique functional principle could provide a unifying

explanation of the full range of these phenomena.

Even though the circuits linking neurons in visual cortex are still a matter of investigation, the nature of their properties suggest that the emergence of nCRF phenomena is a consequence of the interplay between different cortical mechanisms (Angelucci et al., 2017) that employ orientation-specific interactions between neurons with spatially separate cRFs. Anatomical studies have established that long-range horizontal connections in V1 have a patchy pattern of origin and termination, link preferentially cortical domains of similar functional properties, such as orientation columns, ocular dominance columns and CO compartments (Gilbert and Wiesel, 1989; Malach et al., 1993; Bosking et al., 1997) and extend up to 8 mm (Gilbert and Wiesel, 1979, 1989). Although the functional specificity of feedback connections from extra-striate cortex is more controversial, some studies (Angelucci et al., 2002; Shmuel et al., 2005) have reported that terminations of V2-V1 feedback projections are also clustered and orientation-specific, providing input from regions that are on average 5 times larger than the cRF. These results make both horizontal and feedback connections well-suited candidates for mediating contextual effects, potentially with different roles for different spatio-temporal integration processes.

Is it possible to interpret the structure of these connections in terms of the purpose they serve?

For building a model of visual information processing from first principles, a crucial observation is that visual scenes are generated by a mixture of elementary causes. Typically, in any given scene, only *few* of these causes are present (Simoncelli and Olshausen, 2001). Hence, for constructing a neural explanation of natural stimuli, sparseness is likely to be a key requirement. Indeed electrophysiological experiments have demonstrated that stimulation of the nCRF increases sparseness in neural activity and decorrelates population responses, in particular under natural viewing conditions (Haider et al., 2010; Vinje and Gallant, 2000; Wolfe et al., 2010). Perhaps the most influential work that linked sparseness to a form of neural coding that could be employed by cortical neurons was the paradigm introduced by Olshausen and Field (1996). After it was shown that sparseness, combined with unsupervised learning using natural images, was sufficient to develop features which resemble receptive fields of primary visual cortex (Olshausen and Field, 1996, 1997; Bell and Sejnowski, 1997; Rehn and Sommer, 2007), a number of extensions have been proposed that have successfully explained many other aspects of visual information processing, such as complex cell properties (Hyvärinen and Hoyer, 2001) and topographic organization (Hyvärinen et al., 2001). Moreover, a form of code based on sparseness has many potential benefits for neural systems, being energy efficient (Niven and Laughlin, 2008), increasing storage capacity in associative memories (Baum et al., 1988; Charles et al., 2014) and making the structure of natural signals explicit and easier to read out at subsequent level of processing (Olshausen and Field, 2004). Particularly noteworthy is the fact that these statistical models can be reformulated as dynamical systems (Rozell et al., 2008), where processing units can be identified with real neurons having a temporal dynamics that can be implemented with various degrees of biophysical plausibility: using local learning rules (Zylberberg et al., 2011), spiking neurons (Hu et al., 2012; Shapero et al., 2013) and even employing distinct classes of inhibitory neurons (King et al., 2013; Zhu and Rozell, 2015). In summary, sparse coding models nicely explain fundamental properties of vision such as classical receptive fields.

But can these models also explain signatures of contextual processing, namely non-classical receptive fields?

Recently, Zhu and Rozell reproduced a variety of key effects such as surround suppression,

cross-orientation facilitation, and stimulus contrast-dependent ncRF modulations (Zhu and Rozell, 2013). In their framework, small localized stimuli are best explained by activating the unit whose input field ('dictionary' vector) best matches the stimulus. If the stimuli grow larger, other units become also activated and compete for representing a stimulus, thus inducing ncRF modulations. This mechanism is similar to Bayesian models in which contextual effects are caused by surround units 'explaining away' the sensory evidence provided to a central unit (Lochmann et al., 2012). The necessary interactions between neural units are mediated by couplings whose strengths are anti-proportional to the overlaps of the units' input fields. However, most of the effects observed in experiments are caused by stimuli extending far beyond the range of the recorded neuron's input fields (Polat et al., 1998; Walker et al., 2000; Mizobe et al., 2001). Hence the mechanism put forward by this model (Zhu and Rozell, 2013) can only be a valid explanation for a small part of these effects, covering situations in which the surround is small and in close proximity to the cRF. This observation raises the important question, how sparse coding models have to be extended to better reflect cortical dynamics and anatomical structure. In particular, such models would have to allow for direct interactions between non-overlapping input fields.

If these models are then learned from natural images, which local and global coupling structures emerge, how do they compare to anatomical findings, and do they still exhibit the expected cRF properties? Can inference and learning dynamics be implemented in a biophysically realistic manner? Are such models capable of providing satisfactory explanations of ncRF phenomena, and what are the underlying mechanisms? And finally, which predictions emerge from modeling and simulation for experimental studies?

In this paper, we address the above questions by building a novel framework to better capture contextual processing within the sparse coding paradigm. In particular, we define a generative model for visual scenes that takes into account spatial correlations in natural images. To perform inference in this model, we derive a biologically inspired dynamics and a lateral connection scheme that can be mapped onto a neural network of populations of neurons in visual cortex. We show that the emerging connectivity structures have similar properties to the recurrent interactions in cortex. Finally, we evaluate the model's ability to predict empirical findings reported in a set of electrophysiological experiments and we show that it replicates several hallmark effects of contextual processing. In summary, our model provides a unifying framework for contextual processing in the visual system proposing a well-defined functional role for horizontal axons.

3.2 Results

3.2.1 Extended generative model

The low-level, pixel representation of a natural image is multidimensional and complex. However, the corresponding scene can often be described by a much smaller number of high-level, spatially extended components such as textures, contours or shapes, which in turn are composed of more elementary, localized features such as oriented lines or grating patches. Standard sparse coding posits that images can be generated from linear combinations of such elementary features. In particular, it proposes that an image patch $\boldsymbol{s} \in \mathbb{R}^M$ can be written as

$$\boldsymbol{s} = \Phi \boldsymbol{a}, \tag{3.1}$$

where the feature vectors $\phi_i \in \mathbb{R}^M$ are arranged in a $M \times N$ matrix Φ often called 'dictionary' and the vector $\boldsymbol{a} \in \mathbb{R}^N$ contains the coefficients with which a particular image can be represented

in feature space. An implicit assumption made by many sparse coding models (e.g. (Olshausen and Field, 1996; Lewicki and Sejnowski, 2000; Karklin and Lewicki, 2003; Rehn and Sommer, 2007)) is that the features are localized and thus have a limited spatial extent. Such assumption is plausible when features are interpreted as the synaptic input fields of cortical neurons, nevertheless it restricts sparse coding models to encoding only *small* patches of much larger images.

For constructing an extended generative model for natural scenes, we want to take into account that the presence of objects in the scene typically induces long-range dependencies among the elementary features – for instance, an oriented edge that belongs to a contour entails the presence of a co-aligned edge in its proximity (Williams and Thornber, 2001; Ernst et al., 2012). We start by considering a discretization of a (potentially large) visual scene. The simplest scenario that still allows to capture dependencies between features situated in different, non-overlapping locations consists in having two adjacent image patches, as the two horizontally aligned square regions indexed by u and v in Fig. 3.1 (A). Next, we assume that the presence of a feature i at one particular location u can be 'explained' by the presence of features j at other locations v via coefficients C_{ij}^{uv}. We illustrate this in Fig. 3.1 (A), where we have highlighted pairs of oriented edge that belong to the same object and that are thus present in both locations of the visual field. With such matrices C^{uv} and C^{vu} that capture the co-occurrence of features in different locations, we can then define the following feature representations

$$\boldsymbol{b}^u = \boldsymbol{a}^u + C^{uv}\boldsymbol{a}^v, \tag{3.2}$$

$$\boldsymbol{b}^v = \boldsymbol{a}^v + C^{vu}\boldsymbol{a}^u. \tag{3.3}$$

Furthermore, we assume a *reversal symmetry* $C_{ij}^{uv} = C_{ji}^{vu}$ for all i, j, which implies $C^{vu} = (C^{uv})^\top$: if the presence of a feature i at location u implies the presence of a feature j at location v, then the presence of a feature j at location v should imply the presence of a feature i at location u to the same extent (Williams and Thornber, 2001). This allows us to drop the indexes u, v from Eqs. (3.2)-(3.3) and write $C^{uv} = C$ and $C^{vu} = C^\top$ with which, finally, we get

$$\boldsymbol{b}^u = \boldsymbol{a}^u + C\boldsymbol{a}^v \tag{3.4}$$

$$\boldsymbol{b}^v = \boldsymbol{a}^v + C^T\boldsymbol{a}^u \tag{3.5}$$

$$\boldsymbol{s}^u = \Phi\boldsymbol{b}^u \tag{3.6}$$

$$\boldsymbol{s}^v = \Phi\boldsymbol{b}^v. \tag{3.7}$$

In what follows, we will interpret the two patches as a 'central' and 'contextual' stimulus. The extension to more than two patches is straightforward and is presented in Appendix A.

Note that such model might be considered as sparse coding with additional wiring constraint. In fact, substituting equations (3.4)-(3.5) into (3.6)-(3.7) and defining

$$\boldsymbol{s} = \begin{bmatrix}\boldsymbol{s}^u \\ \boldsymbol{s}^v\end{bmatrix}, \boldsymbol{a} = \begin{bmatrix}\boldsymbol{a}^u \\ \boldsymbol{a}^v\end{bmatrix} \text{ and } F = \begin{bmatrix}\Phi & \Phi C \\ \Phi C^\top & \Phi\end{bmatrix}$$

yields the classic linear mixture models used to investigate sparse coding of natural scenes (Olshausen and Field, 1997; Hoyer, 2002)

$$\boldsymbol{s} = F\boldsymbol{a}. \tag{3.8}$$

In fact, for $C = 0$, Eq. (3.8) becomes exactly equivalent to the standard sparse coding model, where image patches would be encoded independently without using the potential benefits of long-range dependencies.

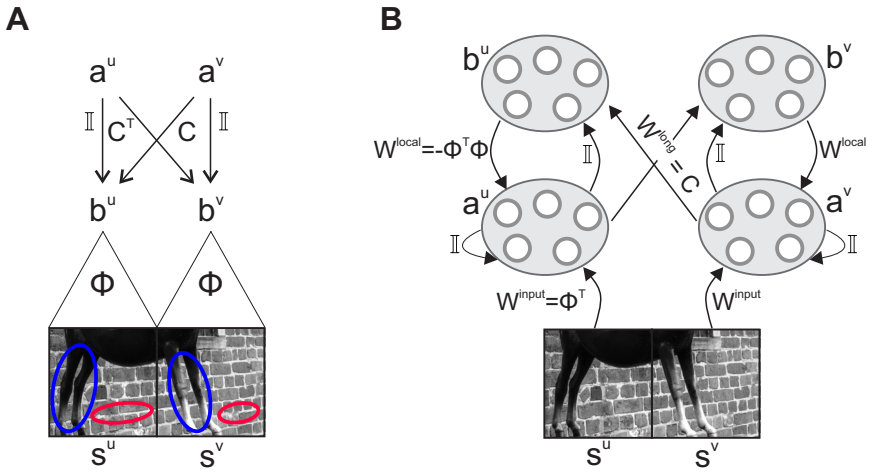


Fig. 3.1: Simplified generative model and neural inference network. (**A**) In a simplified model, we consider visual scenes composed of two horizontally aligned, separate image patches which are encoded by their sparse representation $\boldsymbol{a}^u$, $\boldsymbol{a}^v$ via local features Φ and non-local dependencies C. The highlighted regions indicate how particular pair of local features may co-occur due to the long-range dependencies induced by spatially extended objects. (**B**) Inference in the simplified generative model can be performed by a neural population dynamics (3.22) whose activities represent the coefficients $\boldsymbol{a}^u$, $\boldsymbol{a}^v$ and $\boldsymbol{b}^u$, $\boldsymbol{b}^v$. The corresponding neural circuit involves feedforward, recurrent, and feedback interactions which are functions of the dictionary Φ and of the long-range dependencies C.

Learning visual features and their long-range dependencies. To fully define the coding model we posit an objective function, used for optimization of the latent variables and the parameters. In our scheme, it allows to learn which fundamental features ϕ_i are best suited to encode an ensemble of images, and to derive a suitable inference scheme for the latent variables $\boldsymbol{a}^u, \boldsymbol{a}^v$ such that they optimally explain a given input image $(\boldsymbol{s}^u, \boldsymbol{s}^v)$ given the constraints. Most importantly, it allows to determine the spatial relations C between pairs of features.

The objective function E consists of four terms. The first two quantify how well the two image patches are represented, by means of computing the quadratic error between the patches and their reconstruction. The third and fourth terms require the representation in the coefficients $\boldsymbol{a}$'s and the matrix C to be sparse, which is crucial for our assumption that only few non-zero coefficients are necessary to represent a complex image $(\boldsymbol{s}^u, \boldsymbol{s}^v)_\mu$ from an ensemble of images $\mu = 1, \ldots, P$. Mathematically, it is defined as

$$E_\mu(\boldsymbol{a}^u, \boldsymbol{a}^v, \Phi, C) = \left\|\boldsymbol{s}^u_\mu - \Phi(\boldsymbol{a}^u + C\boldsymbol{a}^v)\right\|_2^2 + \left\|\boldsymbol{s}^v_\mu - \Phi(\boldsymbol{a}^v + C^\top \boldsymbol{a}^u)\right\|_2^2 + \lambda_a \left(\|\boldsymbol{a}^u\|_1 + \|\boldsymbol{a}^v\|_1\right) + \lambda_C \|C\|_2^2 \,. \tag{3.9}$$

The parameters λ_a and λ_C are sparseness constants, with larger values implying sparser representations. To obtain the matrices Φ and C we used a gradient descent with respect to $\boldsymbol{a}^u$, $\boldsymbol{a}^v$, Φ and C on the objective function defined by Eq. (3.9). As image patches $(\boldsymbol{s}^u, \boldsymbol{s}^v)$ we used pairs of neighboring quadratic patches (aligned either horizontally or vertically) extracted from natural images (McGill data set (Olmos and Kingdom, 2004)) after applying a whitening procedure as described in (Olshausen and Field, 1997). Our optimization scheme consisted of two alternating steps: First, we performed inference for an ensemble of image patches by

iterating, for each image μ,

$$\boldsymbol{a}_\mu^{\text{new}} = \boldsymbol{a}_\mu^{\text{old}} - \eta_a \frac{\partial E_\mu}{\partial \boldsymbol{a}} \tag{3.10}$$

until convergence to a steady state while holding Φ and C fixed. Then, we updated Φ or C by computing

$$\Phi^{\text{new}} = \Phi^{\text{old}} - \eta_\Phi \left\langle \frac{\partial E}{\partial \Phi} \right\rangle_\mu \tag{3.11}$$

or

$$C^{\text{new}} = C^{\text{old}} - \eta_C \left\langle \frac{\partial E}{\partial C} \right\rangle_\mu \tag{3.12}$$

with learning rates η_Φ and η_C, respectively. Angle brackets $\langle \cdots \rangle_\mu$ denotes the average over the image ensemble while keeping the $\boldsymbol{a}$'s at the steady states (for details, see Section 3.4.1). This learning schedule reflects the usual assumption that inference and learning take place at different time scales. For increasing computational efficiency, we performed optimization in two phases. First, using only Eqs. (3.10) and (3.11), we learned the dictionary Φ assuming $C = 0$, and second, using only Eqs. (3.10) and (3.12), we obtained the long-range dependencies C while holding Φ fixed.

3.2.2 Inference with a biologically plausible dynamics

While in theory inference and learning can be realized by the general optimization scheme presented above, in the brain inference needs to respect neurobiological constraints. In what follows, we derive a dynamics where the mixture coefficients $\boldsymbol{a}^u$, $\boldsymbol{a}^v$ and $\boldsymbol{b}^u$, $\boldsymbol{b}^v$ are activities of populations of neurons which we hypothesize to realize the necessary computations in cortical hyper-columns connected by local and long-range recurrent interactions (see Fig. 3.1 (B)). Hereby we require populations to have direct access only to 'local' image information, conveyed by their synaptic input fields.

For inference, we assume the quantities Φ and C to be given and we associate each feature i to one neural population having an internal state (e.g. an average membrane potential) and an activation level (i.e., its average firing rate). Following the approach of Rozell et al. (Rozell et al., 2008), we define the population activities $\boldsymbol{a}^X = (a_j^X)_{j=1,\ldots,N}$ as the thresholded values of the internal states $\boldsymbol{h}^X = (h_j^X)_{j=1,\ldots,N}$ by setting

$$\boldsymbol{a}^X = [\boldsymbol{h}^X - \lambda_a]^+, \text{ for } X = u, v, \tag{3.13}$$

using the sparseness constant λ_a as a threshold, and we let $\boldsymbol{h}^X$ evolve according to

$$\tau_h \dot{\boldsymbol{h}}^X = -\frac{\partial E}{\partial \boldsymbol{a}^X}. \tag{3.14}$$

The linear threshold operation ensures the positivity of $\boldsymbol{a}$, which is a necessary requirement for a neural output. Writing (3.14) explicitly leads to

$$\tau_h \dot{\boldsymbol{h}}^u = -\boldsymbol{h}^u + \Phi^\top \boldsymbol{s}^u - (\Phi^\top \Phi - \mathbb{I}_N + C\Phi^\top \Phi C^\top)\boldsymbol{a}^u - (C\Phi^\top \Phi + \Phi^\top \Phi C)\boldsymbol{a}^v + C\Phi^\top \boldsymbol{s}^v \tag{3.15}$$

$$\tau_h \dot{\boldsymbol{h}}^v = -\boldsymbol{h}^v + \Phi^\top \boldsymbol{s}^v - (\Phi^\top \Phi - \mathbb{I}_N + C^\top \Phi^\top \Phi C)\boldsymbol{a}^v - (C^\top \Phi^\top \Phi + \Phi^\top \Phi C^\top)\boldsymbol{a}^u + C^\top \Phi^\top \boldsymbol{s}^u. \tag{3.16}$$

Interpreting these equations in a neural context reveals one problem: The dynamics of the populations at location u explicitly depends on the 'stimulus' (image patch) at location v – and vice versa (last terms on the r.h.s of Eq. (3.15) and (3.16)). This dependency violates our assumption of populations having access to only local image information. One way to get rid of this dependence is to approximate the input by its reconstruction suggested by the generative model, that is $\boldsymbol{s}^u = \Phi(\boldsymbol{a}^u + C\boldsymbol{a}^v)$ and $\boldsymbol{s}^v = \Phi(\boldsymbol{a}^v + C^T\boldsymbol{a}^u)$, which leads to

$$\tau_h \dot{\boldsymbol{h}}^u = -\boldsymbol{h}^u + \Phi^\top \boldsymbol{s}^u - (\Phi^\top \Phi - \mathbb{I}_N)\boldsymbol{a}^u - \Phi^\top \Phi C \boldsymbol{a}^v \tag{3.17}$$

$$\tau_h \dot{\boldsymbol{h}}^v = -\boldsymbol{h}^v + \Phi^\top \boldsymbol{s}^v - (\Phi^\top \Phi - \mathbb{I}_N)\boldsymbol{a}^v - \Phi^\top \Phi C^\top \boldsymbol{a}^u. \tag{3.18}$$

These two equations can be further simplified by extending the dynamical reformulation to include the coefficients $\boldsymbol{b}$ using Eqs. (3.4) and (3.5). For this, we define another set of internal variables $\boldsymbol{k}^X$ satisfying

$$\boldsymbol{b}^X = [\boldsymbol{k}^X]^+ \tag{3.19}$$

and let them evolve according to a similar relaxation equation (i.e. leaky integration):

$$\tau_k \dot{\boldsymbol{k}}^u = -\boldsymbol{k}^u + \boldsymbol{a}^u + C\boldsymbol{a}^v \tag{3.20}$$

$$\tau_k \dot{\boldsymbol{k}}^v = -\boldsymbol{k}^v + \boldsymbol{a}^v + C^\top \boldsymbol{a}^u. \tag{3.21}$$

The final model is thus given by the following four differential equations

$$\begin{cases} \tau_h \dot{\boldsymbol{h}}^u = -\boldsymbol{h}^u + \Phi^\top \boldsymbol{s}^u - \Phi^\top \Phi \boldsymbol{b}^u + \boldsymbol{a}^u & = -\boldsymbol{h}^u + W^{\text{input}} \boldsymbol{s}^u + W^{\text{local}} \boldsymbol{b}^u + \boldsymbol{a}^u \\ \tau_h \dot{\boldsymbol{h}}^v = -\boldsymbol{h}^v + \Phi^\top \boldsymbol{s}^v - \Phi^\top \Phi \boldsymbol{b}^v + \boldsymbol{a}^v & = -\boldsymbol{h}^v + W^{\text{input}} \boldsymbol{s}^v + W^{\text{local}} \boldsymbol{b}^v + \boldsymbol{a}^v \\ \tau_k \dot{\boldsymbol{k}}^u = -\boldsymbol{k}^u + \boldsymbol{a}^u + C\boldsymbol{a}^v & = -\boldsymbol{k}^u + \boldsymbol{a}^u + W^{\text{long}} \boldsymbol{a}^v \\ \tau_k \dot{\boldsymbol{k}}^v = -\boldsymbol{k}^v + \boldsymbol{a}^v + C^\top \boldsymbol{a}^u & = -\boldsymbol{k}^v + \boldsymbol{a}^v + \left(W^{\text{long}}\right)^\top \boldsymbol{a}^u \end{cases} \tag{3.22}$$

and by the linear threshold operations of Eqs. (3.13) and (3.19).

This temporal dynamics can be implemented in a network of four neural populations organized in two cortical columns (Fig. 3.1 (B)). Specifically, populations $\boldsymbol{a}^u$ and $\boldsymbol{a}^v$ in the two columns receive *feed-forward* input $W^{\text{input}} = \Phi^\top$ from two different locations in the visual field. The input is then processed by a set of recurrent *local* connections that couple population $\boldsymbol{a}^u$ to $\boldsymbol{b}^u$ and $\boldsymbol{a}^v$ to $\boldsymbol{b}^v$ within the same column (matrices $\mathbb{I}$ and $W^{\text{local}} = -\Phi^\top \Phi$). The two populations $\boldsymbol{b}^u$ and $\boldsymbol{b}^v$ are also targets of *long-range* connections $W^{\text{long}} = C$ and $(W^{\text{long}})^\top$ originating from populations $\boldsymbol{a}^v$ and $\boldsymbol{a}^u$ in the neighboring column, respectively. For example, the two populations $\boldsymbol{a}$ and $\boldsymbol{b}$ inside a column could be interpreted as neural ensembles located in different cortical layers, or alternatively as two subpopulations in the same layer, but with different connection topologies. Note that the term 'long-range' not necessarily relates to long-ranging horizontal interactions – different anatomical interpretations are possible, and we will speculate on two alternative explanations in the Discussion.

The computation performed within single columns implements a competition based on tuned inhibition between units that code for similar features – which is a typical characteristics of sparse coding models – and it produces a sparse representation of the incoming stimulus. The interactions conveyed by horizontal connections between columns can induce modulatory effects on such a representation. All these connection patterns are completely determined by the matrices Φ and C.

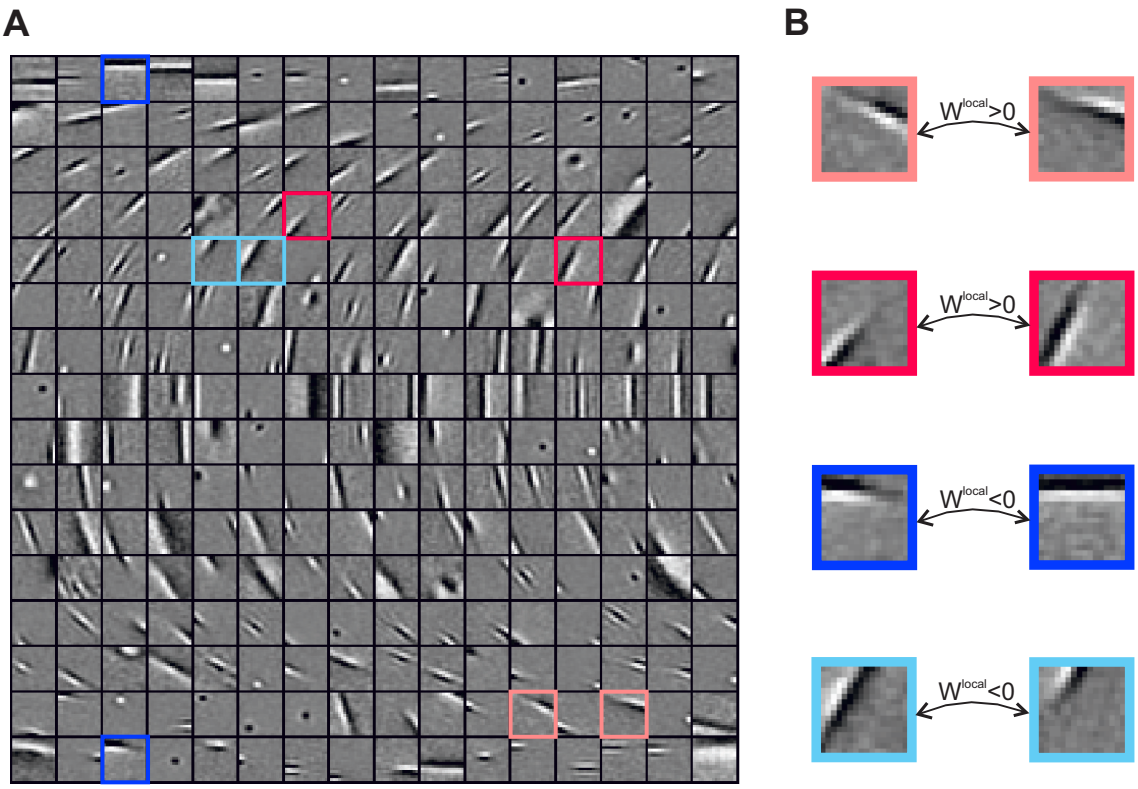


Fig. 3.2: Dictionary Φ and local connections. (**A**) Feature vectors learned by training the model on natural images resemble localized Gabor filters. Features are ordered according to their orientation, which was estimated by fitting a Gabor function. Only a subset of the total set of $N = 1024$ dictionary elements is shown. (**B**) Units with overlapping input fields have strong short-range connections. The sign of the coupling is determined by the arrangement of on/off regions of the input fields: opposite phases correspond to excitatory connections (red) and matching phases to inhibitory connections (blue).

Since the representations $\boldsymbol{a}$ and $\boldsymbol{b}$ can contain both positive and negative entries, each original unit can be realized by two neural populations which we will term 'ON' and 'OFF' units. Hereby ON-units represent positive activations of the original units, while OFF-units represent negative activations of the original units through positive neural activities. Accordingly, OFF-units are assigned the same shape of the synaptic input field, but with opposite polarity. With this necessary extension, Eq. (3.14) implies that $\boldsymbol{a}$ will minimize the energy function E: even though the dynamics does not follow the gradient along the direction of its steepest slope, it still performs a gradient descent, since $\boldsymbol{a}$ is a monotonously increasing function of $\boldsymbol{h}$. We note here that, despite converging to the same fixed point, the dynamics defined by Eq. (3.22) is not equivalent to performing standard gradient descent as in Eq. (3.10) (read Discussion for a more complete explanation).

3.2.3 Connection patterns and topographies

The link between the formal generative model and its realization as a cortical network allows to interpret Φ and C (shown in Fig. 3.2 and 3.4) in terms of the connection matrices W^{input}, W^{local} and W^{long}.

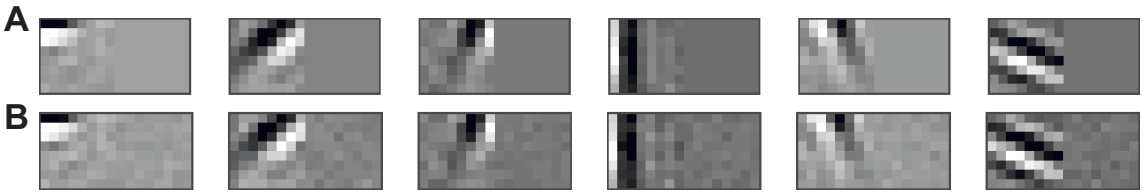


Fig. 3.3: Receptive field mapping. (**A**) Six randomly chosen examples of learned feature vectors. (**B**) Receptive fields corresponding to the feature vectors shown in (**A**), obtained by mapping with spot noise.

After convergence of the training procedure (see Section 3.4.1) our model produces feature vectors that resemble Gabor filters (Fig. 3.2 (A)), having spatial properties similar to those of V1 receptive fields. This result is a consequence of the sparseness constraint and does not come as a surprise, since it was obtained in a number of studies before (Olshausen and Field, 1996; Bell and Sejnowski, 1997; Rehn and Sommer, 2007), but verifies that our extended framework produces meaningful results by being able to learn similar features. The variety of the dictionary elements is represented in Fig. 3.2 (A) and contains examples of localized and oriented Gabor-like patches, concentric shapes, and structures with multiple, irregularly shaped subfields. Each of the dictionary elements represents the synaptic input field of a neural unit and typically shows up as its classical RF when mapped with localized random stimuli through a reverse correlation procedure, as shown in Fig. 3.3. For further analysis, we extracted parameters that characterize the cell's tuning properties – namely its orientation preference, spatial frequency preference, RF center and size – by fitting a Gabor filter to each feature vector (see Section 3.4.1). Typically, all feature vectors taken together build a complete representation for all orientations (and other stimulus features), thus the columns indicated in Fig. 3.1**B** are similar to orientation hypercolumns found in primary visual cortex (Hubel and Wiesel, 1974). The distribution of orientation preferences exhibits a bias for cardinal orientations as observed in physiological studies (Wang et al., 2003).

As previously mentioned, short-range interactions are specified by the dictionary matrix through the equation $W^{\text{local}} = -\Phi^{\top}\Phi$. This implies that the absolute strength with which two units are locally connected is proportional to how closely their respective input fields match. In particular, as it is illustrated in Fig. 3.2 (B), units with similar orientation preference and opposite phase are excitatorily connected, while units with similar orientation and similar phase are inhibitorily connected. Support for such like-to-like suppression can be found in a recent experiment (Chettih and Harvey, 2019), where optogenetic stimulation of mice V1 revealed a prominent inhibitory influence between neurons with similar tuning, suggesting that feature competition is indeed implied in sensory coding.

Together with the dictionary, we also learn the long-range feature dependencies C (Fig. 3.4 shows results relative to the horizontal configuration). To investigate which pattern of connections is induced, we computed the average absolute connection strength $\langle |W^{\text{long}}(\theta_{\text{post}}, \theta_{\text{pre}})| \rangle$ as a function of the orientation preferences θ_{post} and θ_{pre} of the units they connect (Fig. 3.4 (A)). The highest absolute connection strengths appear along the diagonal, indicating that pairs of units with similarly oriented input fields tend to be more strongly connected via long-range interactions. The distribution contains another structure, although more faint, located along the anti-diagonal, indicating that pairs of units whose orientations sum up to 0 degrees are also strongly connected. Particular examples of units that have strong long-range coupling are shown in Fig. 3.4 (A).

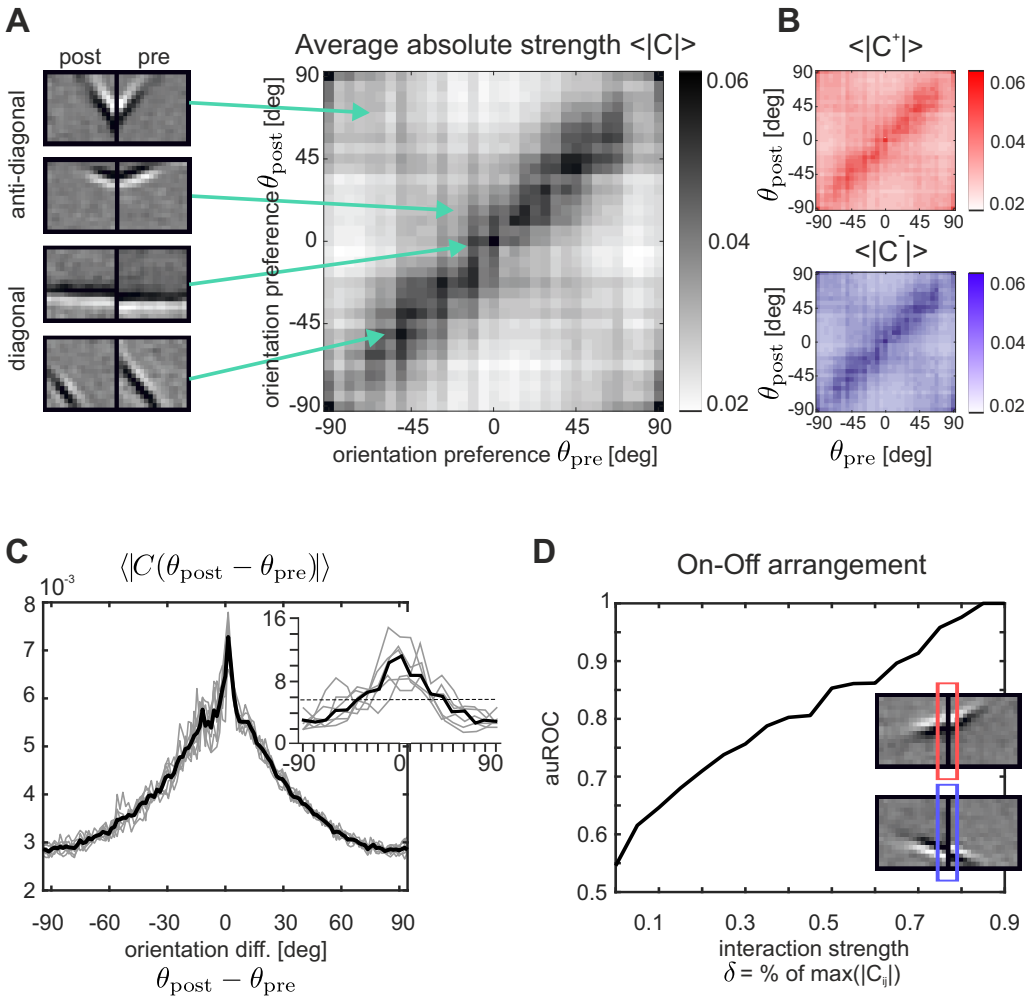


Fig. 3.4: Correlation matrix C and long-range interactions. (**A**) Average absolute strength of long-range connections $W^{\text{long}} = C$ as a function of the orientation preferences of the pre- and postsynaptic units. Each point in the graph represents a connection from a unit responsive to the right portion of the visual field to a unit responsive to the left portion of the visual field (see four examples on the left). (**B**) Average absolute strength of excitatory (top, red color scale) and inhibitory (bottom, blue color scale) long-range connections as a function of the pre- and postsynaptic orientation preferences as in (**A**). (**C**) Average absolute strength of long-range connections as a function of the difference in orientation preference of the connected units. For comparison, data from the primary visual cortex of tree shrews are shown in the inset. The graph displays the percentage of boutons contacting postsynaptic sites that differ in orientation preference by a specified amount from the presynaptic injection site of a biocytin tracer. Individual cases are shown in gray and the median is shown in black. The dashed line reflects the percentage of boutons expected in each orientation difference bin if the boutons were distributed evenly over the map of orientation preference (redrawn from (Bosking et al., 1997)). (**D**) Long-range interactions between units having positive correlations between the adjacent borders of their synaptic input fields tend to be excitatory (red frame in upper input fields example), while units having negative correlations tend to be inhibitory. This effect increases with increasing absolute coupling strengths $|C_{ij}|$, as indicated by the area under the ROC curve (auROC) computed from the corresponding correlation distributions for positive and negative connections.

This result is consistent with anatomical measurements taken in primary visual cortex of mammals. Several experiments (Gilbert and Wiesel, 1989; Malach et al., 1993; Weliky et al., 1995; Yoshioka et al., 1996; Bosking et al., 1997) report that horizontal long-range connections in V1 show a 'patchy' pattern of origin and termination, linking preferentially cortical domains responding to similar features. We quantified such a tendency in our model by computing the average connection strength as a function of the orientation preference difference $\Delta\theta = \theta_{\text{post}} - \theta_{\text{pre}}$ between pre- and post-synaptic cell. The corresponding graph is shown in Fig. 3.4 (C), and a similar distribution obtained from anatomical measurements is reported for comparison in the inset.

In three shrew (Bosking et al., 1997), cat (Schmidt et al., 1997) and monkeys (Sincich and Blasdel, 2001), it has been shown that long-range connections between neurons of similar orientation selectivity exist primarily for neurons that are retinotopically aligned along the direction of their cells' preferences. We computed average absolute coupling strength between populations with aligned cRFs (i.e., 0 ± 15 degrees), and between populations with parallel cRFs (i.e., 90 ± 15 degrees), revealing that aligned couplings were indeed 26% percent stronger on average.

When splitting long-range interactions into negative and positive weights, we do not find any significant difference between their dependency on pre- and postsynaptic orientation preference (Fig. 3.4 (B)). However, a different pattern emerges when we take the polarities or phases of the synaptic input fields into account: For this purpose we measured the correlation ρ between the right border of the left input field, and the left border of the right input field (colored frames in inset of Fig. 3.4 (D)), which are adjacent in visual space. Excitatory connections tend to exhibit positive correlations, while inhibitory connections tend to exhibit negative correlations. The stronger the couplings, the more pronounced this effect becomes. To quantify this effect, we compared the distribution of correlations between elements linked by positive couplings larger than δ with the distribution of correlations between elements linked by negative couplings smaller than $-\delta$, namely

$$p(\rho \,|\, W_{ij}^{\text{long}} > \delta) \text{ and } p(\rho \,|\, W_{ij}^{\text{long}} < -\delta), \tag{3.23}$$

by computing a receiver-operator characteristics ROC. Consistently, we find that separability as quantified by the area under ROC (auROC) increases with δ (Fig. 3.4 (D)). This effect is opposite to what we have (by construction) for the short-range connections: while units with similar cRFs *within a column* compete with each other, units with similar cRFs *across two columns* facilitate each other.

3.2.4 Contextual effects

With the input fields (dictionary) and the long-range interactions obtained from a representative ensemble of natural images, the connectivity of the network represented in Fig. 3.1 is completely specified. We can then subject the model to arbitrary stimulus configurations and investigate how well the dynamics described by Eqs. (3.13), (3.19) and (3.22) predicts key effects exhibited by real neurons when processing contextual visual stimuli, and whether it can offer a coherent explanation to experimentally established context effects. For this purpose, we first selected units that were well driven and well tuned to the orientation θ_c of small patches s_c of drifting sinusoidal gratings positioned at the center $\boldsymbol{r}^u$ of the left input region (cf. Fig. 3.2 (B)),

$$s_c(\boldsymbol{r}, t) = k_c \gamma_c(\boldsymbol{r}) \sin\left(\omega_c(\boldsymbol{r} - \boldsymbol{r}^u)\boldsymbol{e}_{\theta_c} + \omega_t t\right), \tag{3.24}$$

$$\gamma_c(\boldsymbol{r}) = \frac{1}{2}\left(1 + \tanh\left(\beta(r_c - |\boldsymbol{r} - \boldsymbol{r}^u|)\right)\right).$$

Here k_c denotes grating contrast, r_c the radius of the patch, ω_c its spatial frequency, ω_t the drifting frequency and β controls the steepness of the transition between stimulus und background. Thereby we mimic the situation in experiments in which typically also time-dependent stimuli are used. Subsequently, these selected units were subjected to contextual stimulation, and the induced modulation by the context quantified.

In the following, we will focus on three exemplary stimulation paradigms in contextual processing, assessing size tuning, orientation-contrast effects, and luminance contrast effects.

Size tuning. Experiments in monkey and cat (Sceniak et al., 1999; Walker et al., 2000) have shown that the stimulation of visual space surrounding the classical receptive field often has a suppressive influence on neurons in V1. Stimuli typically used to reveal this effect consist of a moving grating or an oscillating Gabor patch having the cell's preferred orientation, and being positioned at the center of its cRF. Recording the neural response while increasing the size of the grating yields the size tuning curve which exhibits two characteristic response patterns (Walker et al., 2000), as indicated in Fig. 3.5 (A): After an initial increase in firing rate with increasing stimulus size, either the cell's response becomes suppressed and firing rate decreases (upper panel), or firing rate increases further and finally saturates (lower panel). In our model we realized a similar stimulation paradigm by using an optimally oriented grating (Eq. (3.24)) and increasing its size r_c. Hence the stimulus first grows towards the border of the input field in which it is centered, and then extends into the neighboring fields. From all selected units, we show the size tuning curves of two exemplary cells in Fig. 3.5 (B), demonstrating that the model can capture both qualitative behaviors known from cortical neurons.

For quantifying the degree of suppression and the extent to which this effect is present at the population level, we computed for all selected units a suppression index (SI) defined as

$$\mathrm{SI} = 1 - a_{\mathrm{full}} / \max_{r_c}(a(r_c)),$$

where a_{full} was the response to a stimulus fully covering the input field. The SI indicates how much, in percentage, the response of a unit at largest stimulus size is reduced with respect to its maximum response, with 0 meaning no suppression and 1 meaning total suppression. The distribution of the SI across all the simulated cells is plotted in Fig. 3.5 (C). For population $\boldsymbol{a}$, we find values comparable to what has been found experimentally: (Walker et al., 2000) reports that 44% of cells had less than 10% suppression and in the model the percentage of cells with SI< 0.1 is 38%. In general, the model shows less suppression (i.e., lower SI values) for population $\boldsymbol{b}$.

Since surround suppression was already observed in sparse coding models without long-range interactions (Zhu and Rozell, 2013), we expect this effect to stem from a combination of local and long-range connections. To quantify their roles in producing surround suppression, we simulated a version of the model without long-range interactions by setting $C = 0$. The resulting distribution of changes in SI is shown in Fig. 3.5 (D) and displays a mean increase of the SI for population $\boldsymbol{a}$ when including long-range connections, indicating that they contribute considerably to suppressive modulation induced by stimuli in the surround. In fact, without long-range interactions the percentage of cells with SI< 0.1 becomes 64%, which is quite far from the experimental result reported above. Conversely, the effect of including long-range connections is predominantly facilitatory for population $\boldsymbol{b}$, leading to a decrease in the observed SI's.

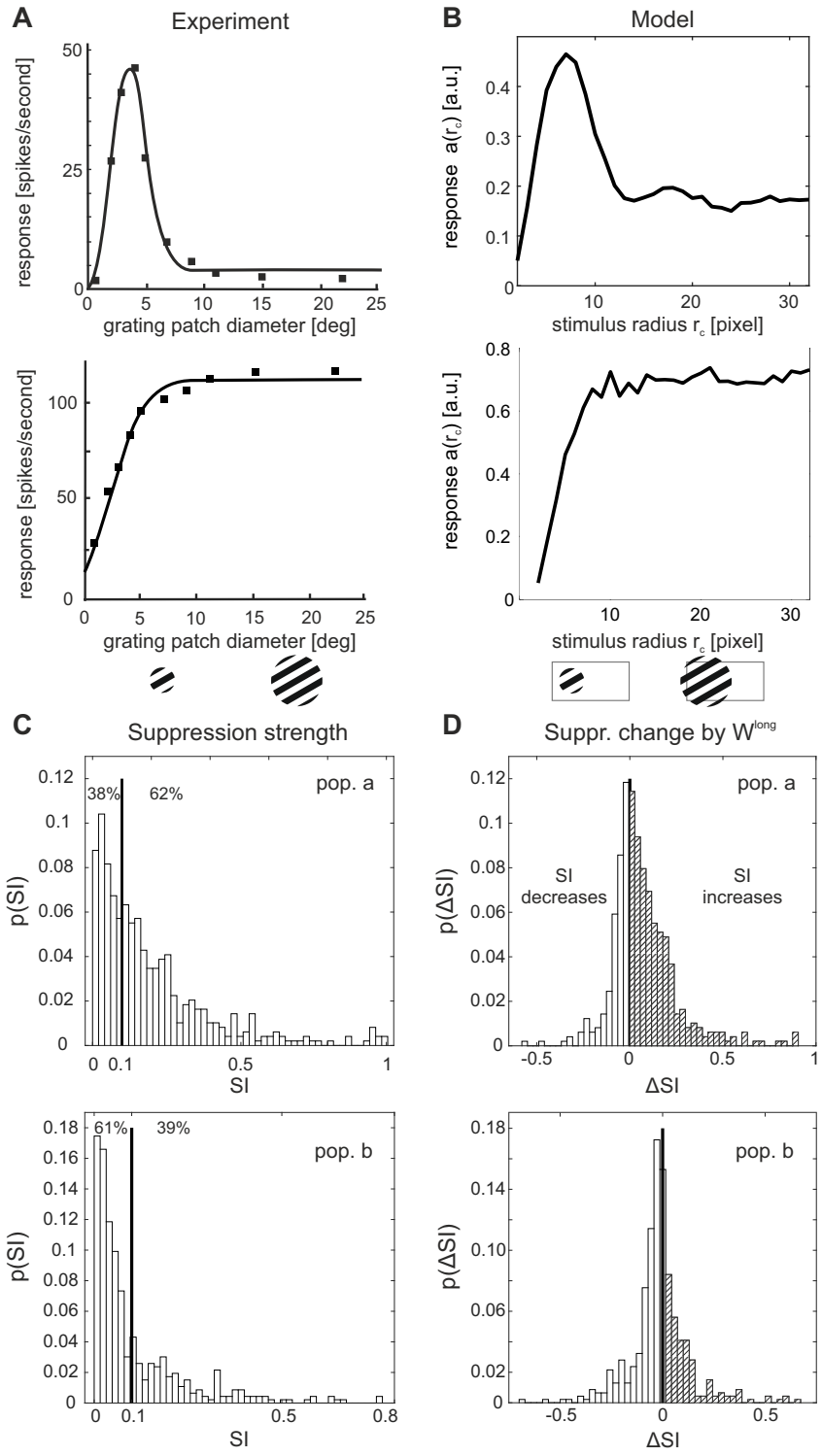


Fig. 3.5: Size tuning and surround suppression. Dependence of neural responses on the size of a circular moving grating presented at the cell's preferred orientation. (**A**) Single-cell size tuning curves cat's V1 exhibiting surround suppression (top) or saturation (bottom). Redrawn from (Walker et al., 2000). (**B**) Size tuning curve of exemplary units in the model showing similar behaviour as in (**A**). (**C**) Distribution of suppression indices SI for the full model with long-range interactions. Values of 0 correspond to no suppression, values of 1 to full suppression. (**D**) Change in SI ($\Delta SI = SI^{\text{with long}} - SI^{\text{without}}$) induced by long-range connections. Enhanced suppression occurs more frequently than facilitation in population **a**, while in population **b** one observes the opposite effect.

Cross-orientation modulation. Contextual processing is often probed by combining a central grating patch inside the cRF with a surround annular grating outside the cRF. For such configurations, the influence of the surround annulus on the response to an optimally oriented center stimulus was found to be orientation selective. When center and surround have the same orientation, the firing rate modulation is mostly suppressive, as we already know from studying size tuning (previous subsection).

If the surround strongly deviates from the orientation of the center, suppression becomes weaker (Levitt and Lund, 1997; Sengpiel et al., 1997; Walker et al., 1999; Cavanaugh et al., 2002b) and in some cases even facilitation with respect to stimulation of the center alone is observed (Sillito et al., 1995; Jones et al., 2002). In particular, one study in cats (Sengpiel et al., 1997) reports three typical response patterns: (I) equal suppression regardless of the orientation of the surround, (II) suppression which decays with increasing difference between the orientations of center and surround, and (III) suppression that is strongest for small differences between orientations of center and surround, and weaker for large orientation differences and orientation differences close to zero. In the literature, the last effect is also termed 'iso-orientation release from suppression' (see Fig. 3.6 (A) for examples).

We realized this experimental paradigm in our model by combining a central grating patch (Eq. (3.24)) with a surround annulus

$$s_a(\boldsymbol{r}, t) = k_a \gamma_a(\boldsymbol{r}) \sin\left(\omega_a (\boldsymbol{r} - \boldsymbol{r}^u)\boldsymbol{e}_{\theta_a} + \omega_t t\right), \tag{3.25}$$

$$\gamma_a(\boldsymbol{r}) = \frac{1}{4}\left(1 + \tanh\left(\beta(|\boldsymbol{r} - \boldsymbol{r}^u| - r_i)\right)\right)\left(1 + \tanh\left(\beta(r_a - |\boldsymbol{r} - \boldsymbol{r}^u|)\right)\right)$$

having orientation θ_a, spatial frequency $\omega_a = \omega_c$, inner radius $r_i = r_c$, outer radius r_a, and grating contrast $k_a = k_c$. For each neural unit we investigated, the center stimulus had an optimal size defined by the radius r_c for which we obtained the maximum response in the unit's size tuning curve. The surround annulus had the same parameters as the center patch and extended from the radius of the center patch to the whole input space (as displayed in Fig. 3.6, stimulus icons in the legends). While the center orientation was held at the unit's preferred orientation, the surround orientation θ_a was systematically varied between 0 and π. For this experiment, we selected all units for which their optimal size was not larger than 21 pixels, to ensure that there was still space for a surround annulus in the restricted input space.

The three distinct behaviors observed in the experiments are qualitatively captured by the model: in Fig. 3.6 (B) (dashed lines) we show the orientation tuning curve of selected units of the model. Adding an annular surround stimulus to an optimally oriented center induces modulations which are mostly suppressive and tuned to the orientation of the surround (Fig. 3.6 (B), solid lines). Cross-orientation modulations are summarized across the investigated model subpopulation in Fig. 3.6 (C, D), where responses of cells exhibiting the same qualitative behavior are averaged together, as in the experiment (cf. panel (A), see Section 3.4.3 for a detailed description of the pooling procedure). We distinguish, from top to bottom, untuned suppression, iso-orientation suppression, and iso-orientation release from suppression.

To assess the contributions of long-range connections to these effects, we repeated the experiment with $C = 0$. The population averages over the same categories of behaviors are overlaid in Fig. 3.6 (C, D) in gray. A comparison between the results of the model with and without C shows that long-range interactions induce two different effects: enhancing responses for large orientation differences for cells with untuned surround suppression, and increasing maximum suppression for cells with tuned surround suppression in population ***a***. In particular, we observe strong facilitatory effects in population ***b***. This difference between the two populations might

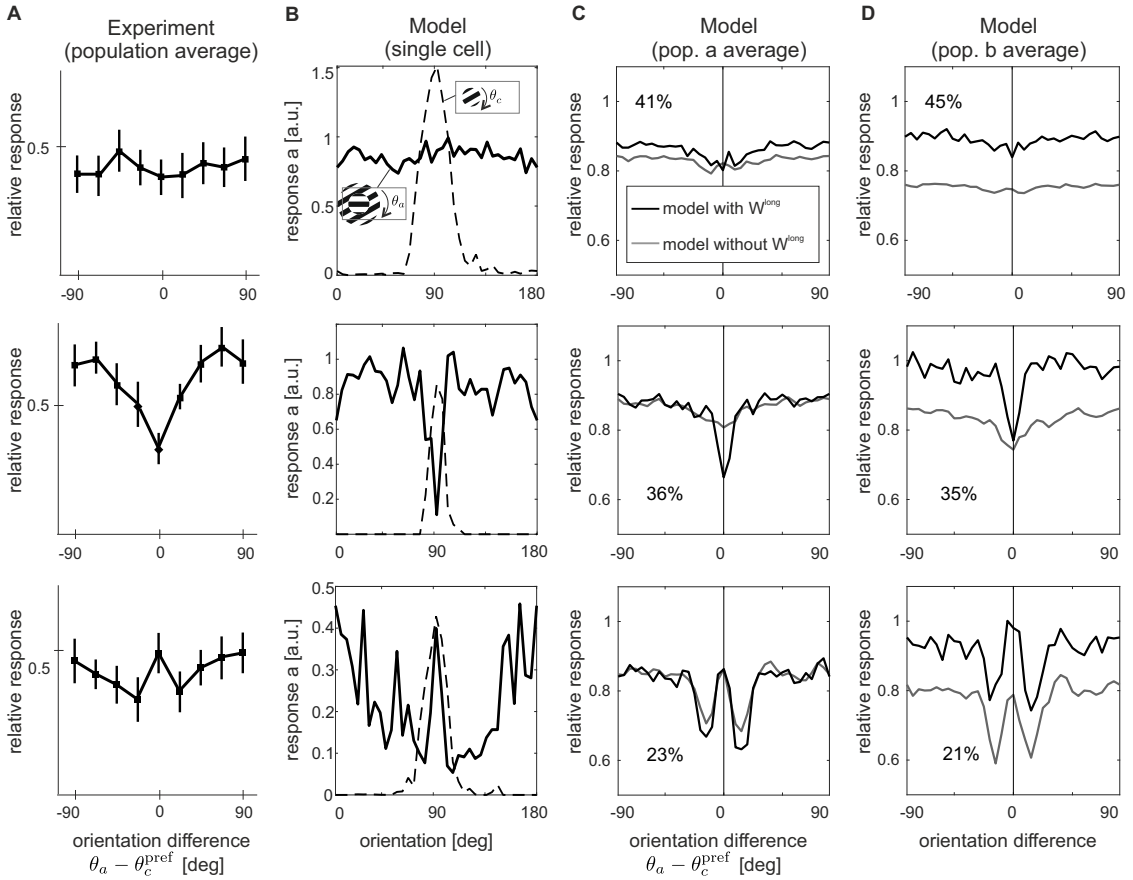


Fig. 3.6: Orientation-contrast modulations. A center stimulus with preferred orientation is combined with an annulus of varying orientations (see icons in column (**B**)). (**A**) In experiments three response patterns are observed, namely, from top to bottom, untuned suppression, iso-orientation suppression and iso-orientation release from suppression (data replotted from (Sengpiel et al., 1997)). The model reproduces these three response patterns both at the single cell level (**B**) and at the population level for a (**C**) and b (**D**). For comparison, orientation tuning for a center-alone stimulus is shown by the dashed line in (**B**). In (**C, D**), the gray lines display orientation-contrast tuning of the same ensembles without long-range interactions. Note that in (**A**) and (**C, D**), responses are shown normalized by the response to the center alone at the preferred orientations of the units. Percentages indicate the proportion of cells that fall in the same orientation-modulation class.

explain an apparent contradiction in experimental data where in a similar orientation contrast tuning paradigm one study exhibited strong facilitation (Sillito et al., 1995), while a different investigation found only moderate release from suppression (Levitt and Lund, 1997).

Luminance-contrast effects. In addition to orientation, also the relative contrast between the brightness of the center and the surround can be varied. In particular, such stimuli often reveal facilitatory effects, which are more frequently observed when the cRF is weakly activated, for example by presentation of a low-contrast visual stimulus. For many cells in V1 ($\approx$ 30%, in (Polat et al., 1998; Chen et al., 2001)), collinear configurations of center-surround stimuli induce both facilitation and suppression. Here the visual contrast of the center stimulus in

comparison to a fixed-contrast surround controls the sign of the modulation, and the point of crossover between suppression and facilitation is related to the cell's contrast threshold (Levitt and Lund, 1997; Mizobe et al., 2001; Polat et al., 1998; Sengpiel et al., 1997; Toth et al., 1996). The characteristics of differential modulation is exemplified in Fig. 3.7 (A) where the contrast response function of a single cell in cat V1 (filled circles) is plotted together with the response of the same cell to the compound stimulus (empty circles). The graph shows that the same surround stimulus can enhance the response to a low-contrast center stimulus and reduce the response to a high-contrast center stimulus.

For obtaining corresponding contrast response curves in our model, we presented each selected unit with a center stimulus of optimal orientation and size of which we varied its contrast k_c (Eq. (3.24)). To mimic the collinear configuration of the compound stimulus, we then placed a surround annulus (Eq. (3.25)) at high contrast $k_a = 1$, iso-oriented with the center patch (see stimulus icons in Fig. 3.7), and again varied the contrast of the center patch. The resulting switch from facilitation to suppression, apparent by the crossing of the two response curves, is well captured by the model and illustrated for an example unit in population ***b*** in Fig. 3.7 (B).

As in previous examples, differential modulation shows considerable variability across recorded cells. In particular, there are V1 neurons which exclusively show suppressive effects, while other neurons exclusively exhibit facilitatory effects. The corresponding statistics is displayed in Fig. 3.7 (C): For each value of contrast that was tested in (Polat et al., 1998), the bars show the proportion of cells that exhibit either facilitation or suppression. In particular, suppression becomes increasingly more common as the contrast of the center stimulus increases. The same analysis applied to our model reveals an identical result (Fig. 3.7 (D)), thus indicating that the model also captures the diversity of behaviors observed in electrophysiology. For population ***b***, the model statistics matches experimental findings also quantitatively. In particular, we observed that the increase in numbers of suppressed cells with increasing center contrast is mainly caused by the long-range connections, since this effect largely disappeared when we set $C = 0$ (horizontal lines in Fig. 3.7 (D)).

3.3 Discussion

The pioneering work of Olshausen and Field (Olshausen and Field, 1996) demonstrated that simple cell responses in primary visual cortex can be understood from the functional requirement that natural images should be represented efficiently by optimally coding an image with sparse activities. Since then, there have been many attempts to derive also other neuronal response properties in visual cortex from first principles. Common to these models is the framework of generative models, where the activities in an area are considered to represent the results of inference in the spirit of Helmholtz (Von Helmholtz, 1962; Doya et al., 2007). Most of these investigations concentrate on local receptive field properties (Olshausen and Field, 1997; Bell and Sejnowski, 1997; Rehn and Sommer, 2007; Hyvärinen and Hoyer, 2001; Hyvärinen et al., 2001). More recently, formal models were introduced that can qualitatively reproduce also several established non-classical receptive field effects (Lochmann et al., 2012; Karklin and Lewicki, 2009; Coen-Cagli et al., 2012, 2015; Zetzsche and Nuding, 2005) and/or predict interactions resembling features of long-ranging horizontal and feedback connections in cortex (Garrigues and Olshausen, 2008; Coen-Cagli et al., 2012).

It is, however, unclear how the networks in cortex might perform the inference these models hypothesize given the neurobiological constraints on anatomy and neuronal dynamics. In this regard, the neural implementation proposed by (Rozell et al., 2008) provided a significant

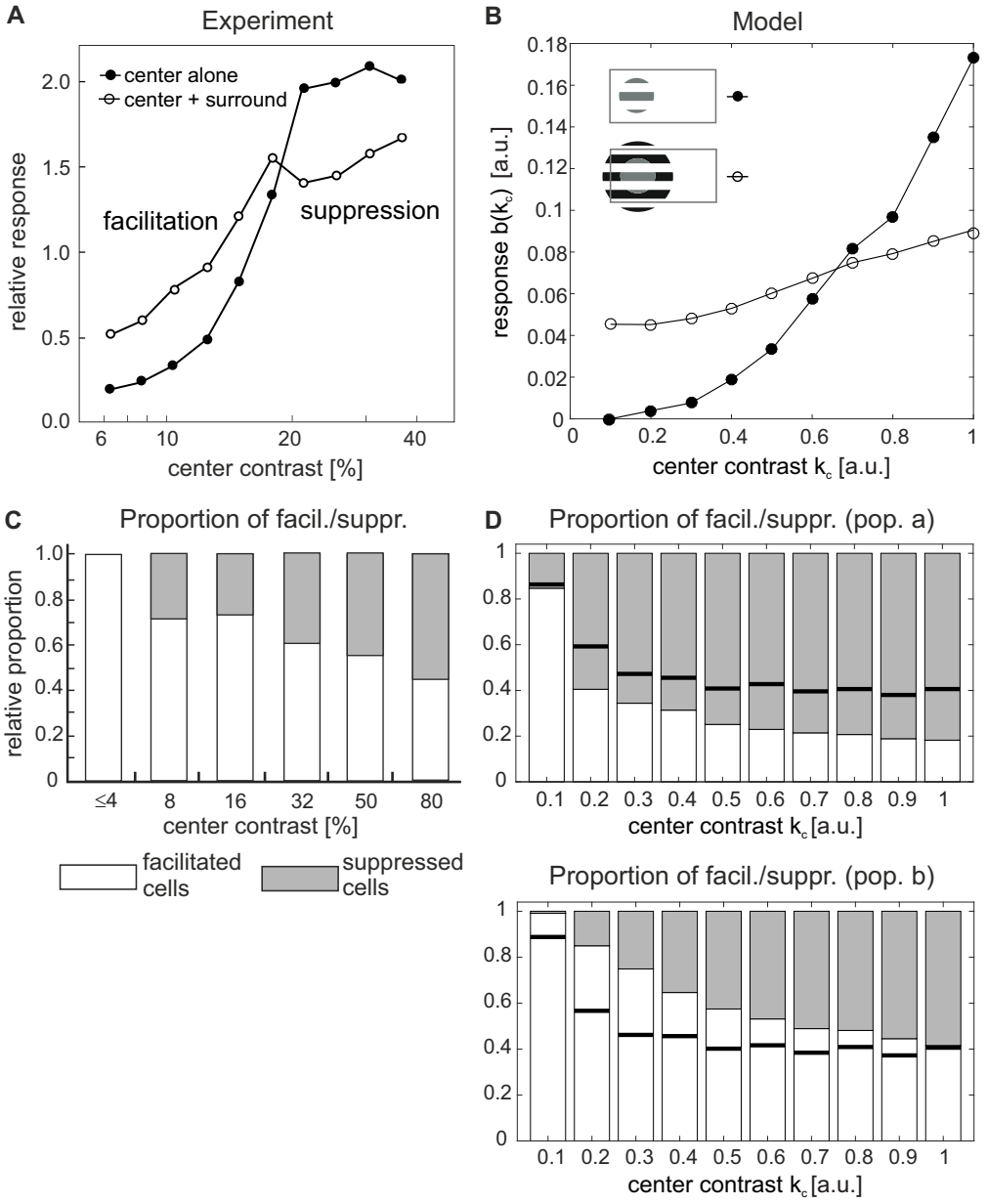


Fig. 3.7: Luminance contrast tuning. (**A, B**) Single-cell responses to a center stimulus of varying contrast without flanking surround stimuli (filled circles) are compared to responses to the same center stimulus combined with high-contrast flanking surround stimuli of the same preferred orientation (open circles) in experiment (**A**) (redrawn from (Polat et al., 1998)) and model (**B**). The stimulus configurations are indicated inside the graphs. (**C, D**) Population statistics, detailing the proportion of cells showing facilitation (light bars) or suppression (gray bars) in dependence on center stimulus contrast. Experimental data in (**C**) is redrawn from (Polat et al., 1998). In the model (**D**), cells were judged to be significantly facilitated (suppressed) if their activation ratio between center-surround and center alone stimulation $\mathsf{b}^{\mathrm{sur}}(\mathsf{k_c})/\mathsf{b}^{\mathrm{cen}}(\mathsf{k_c})$ at contrast $\mathsf{k_c}$ was larger than $1+\varepsilon$ (smaller than $1-\varepsilon$), with $\varepsilon = 0.01$. Solid black lines indicate proportion of cells showing facilitation without long-range interactions. The top plot in (**D**) shows the statistics for population a and the bottom plot for populations b.

advance, since it can explain a range of contextual effects (Zhu and Rozell, 2013) with a neural population dynamics that requires only synaptic summation and can also be extended to obey Dale's law (Zhu and Rozell, 2015). But this model still presents a fundamental, conceptual difference to visual cortex: there are no interactions between neurons with non-overlapping input fields and thus the model can not account for the long-range modulatory influences from far outside the classical receptive field.

Here we propose a generative model for sparse coding of spatially extended visual scenes that includes long-range dependencies between local patches in natural images. An essential ingredient is the inclusion of plausible neural constraints by limiting the spatial extent of elementary visual features, thus mirroring the anatomical restrictions of neural input fields in primary visual cortex.

3.3.1 Relations to standard sparse coding

As it becomes evident rearranging the equations that define the generative model (Eq. (3.8)), our model offers an implementation of sparse coding that allows to encode spatially extended visual scenes. Although it might be tempting to consider it simply as a 'scaled-up' version of (Rozell et al., 2008), we argue that this is indeed not the case. To demonstrate our reasoning, we consider the example of encoding a long horizontal bar. While a scaled-up-sparse-coding would have a specialized long horizontal feature to explain the stimulus (i.e. the sparsest representation), our model, by constraining the features to have a limited size, would require two separate horizontally aligned features to coactively form a representation of the stimulus; such collaborations between neighboring neurons are enforced by long-range connections.

3.3.2 Connection structures

By optimizing model parameters via gradient descent it is possible to determine all connections in the network e.g. from the statistics of natural images. Synaptic input fields Φ resemble classical receptive fields of V1 neurons (Fig. 3.2 (A)). The structure of C turns out to have similar characteristics as the anatomy of recurrent connections in visual cortex, exhibiting a preference to link neurons with similar orientation preferences via long-ranging horizontal axons (Kaschube, 2014; Schmidt et al., 1997; Gilbert and Wiesel, 1989) or via patchy feedback projections (Angelucci et al., 2002; Shmuel et al., 2005). Furthermore, we find a bias for collinear configurations being more strongly connected than parallel configurations, matching the observed elongation of cortical connection patterns along the axis of collinear configurations in the visual field in three shrew (Bosking et al., 1997), cat (Schmidt et al., 1997) and monkeys (Sincich and Blasdel, 2001). These connection properties reflect regularities of the visual environment such as the edge co-occurrence observed in natural images (Geisler et al., 2001).

The role of long-range connections in context integration was investigated also in a recent work (Iyer and Mihalas, 2017). Here the authors assume a neural code in which the firing rate of a neuron selective for a particular feature at a particular location is related to the probability of that feature to be present in an image, and influenced by the probability of other features being present in surrounding locations. In an analogous way as in our model, they assume that the only information a neuron tuned to a specific location in the visual field has about the stimulus context at neighboring locations comes from the neurons that are tuned to those neighboring locations (limited extent of the visual input). Thus, the lateral coupling scheme they obtain is also in good agreement with that observed in V1. Those connections are beneficial in increasing coding accuracy under the influence of noise, but the authors did not critically test their model with contextual stimulus configurations. Since their network does not implement competition,

we expect their model to exhibit surround enhancement for co-aligned stimulus configurations, rather than the experimentally observed suppressive effects.

3.3.3 Learning rules

To be a completely realistic model, still many details are missing. For example, the question of whether connectivity can be learned using realistic plasticity rules remains open. Currently our learning rules (3.11) and (3.12) require the change in single synapses to rely on information from *all* the neurons in the network. Moreover, the analytically derived formula for W^{local} (Eq. (3.22)) implies a pretty tight relation between the short-range interactions and the feed-forward weights and it is not clear which synaptic mechanisms could achieve it in parallel. The local plasticity rules used in (Zylberberg et al., 2011) solved these issues in the context of the standard formulation of sparse coding, but it is not clear if a similar approach could be used to derive a learning rule also for W^{long}.

Finally, our model violates Dale's law, postulating direct inhibitory connections between excitatory cells (for both short- and long-range interactions). In the context of standard sparse coding, some work has been done to improve biological plausibility by implementing inhibition in a separate sub-population of neurons, both in spiking networks (King et al., 2013) and dynamical systems (Zhu and Rozell, 2015). While the first model (King et al., 2013) consists in adding a second population of inhibitory units and then learning separately three sets of weights ($E-I$, $I-E$ and $I-I$), the second (Zhu and Rozell, 2015) relies on a low-rank decomposition of the recurrent connectivity matrix into positive and negative interactions. Both approaches were able to learn a sparse representation code and to develop Gabor-like input fields (notably, using the same E/I ratio observed in visual cortex). However, generalizing either one of them to our extended model might not be straightforward.

3.3.4 Neural dynamics

Inference in the presented model is realized by a biologically realistic dynamics in a network of neural populations that are linked by short- and long-range connections. This implementation of a dynamics is close to the approach of Rozell (Rozell et al., 2008) but additionally includes long-range interactions between units with non-overlapping input fields. Most importantly, the constraint that only local visual information is available to the units receiving direct input from the visual field implies, and predicts, that inference is performed by *two* separate neural populations with activities $\boldsymbol{a}$ and $\boldsymbol{b}$ and different connection structures.

It is worth to speculate about a direct relation to the particular properties of neurons and anatomical structures found in different layers and between areas of visual cortex: Physiological studies distinguish between the near ($<$ 2.5 degrees) and far surround ($>$ 2.5 degrees) in contextual modulation (Angelucci et al., 2017). Taking into account the spread of long-range horizontal axons within V1, which is less than about three degrees in visual space (Angelucci et al., 2002), it seems likely that near surround effects are predominantly caused by horizontal interactions, while far surround effects are rather explained by feedback from higher visual areas. Assuming that one input patch in the model spans across 3 degrees in visual space, which is not implausible given the spatial extent of Gabor-like input fields shown in Fig. 3.2**A** (up to 1 degree in cortex), we would therefore identify 'local' interactions $W^{\text{local}} = -\Phi^{\mathrm{T}}\Phi$ with horizontal axons within V1, while 'long-range' interactions $W^{\text{long}} = C$ would be mediated by the combination of feedforward and feedback connections between visual cortical areas. A possible circuit diagram emerging from this paradigm is depicted in the scheme in Fig. 3.8 (B).

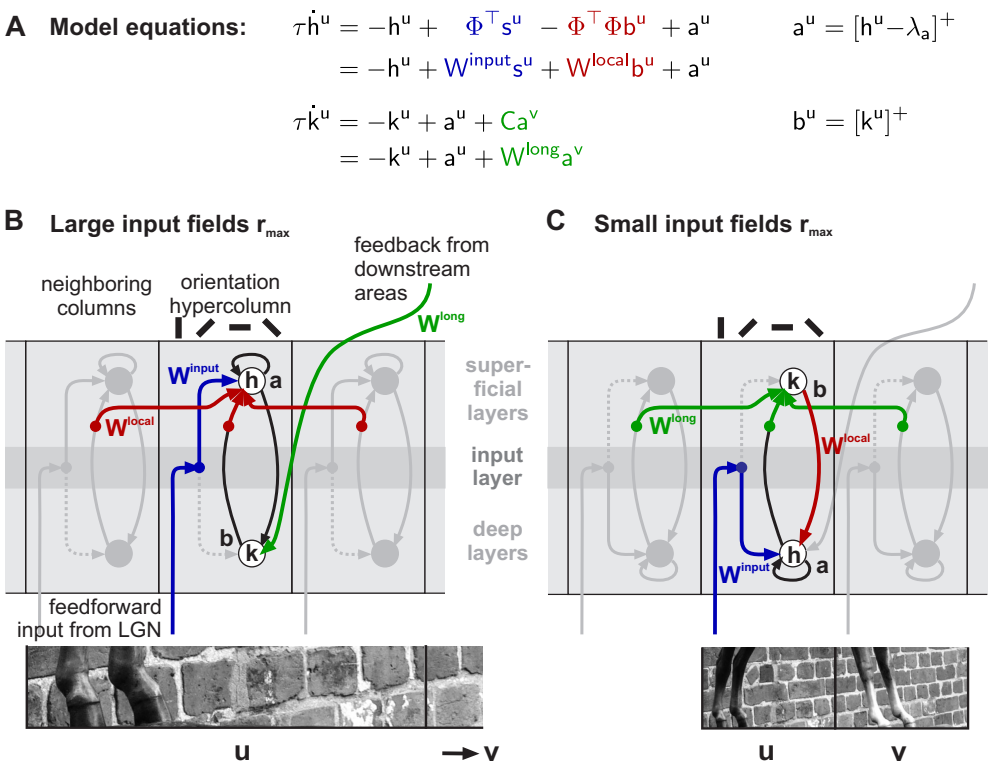


Fig. 3.8: Putative neural circuits performing inference in visual cortex. (**A**) Equations that define the network dynamics. (**B, C**) Depending on the assumed spatial scale of input fields in the generative model, one distinguishes between cortical circuits where 'long-range' interactions W^{long} would be mediated by recurrent loops between different cortical layers and 'local' interactions W^{local} by long-ranging horizontal axons within primary visual cortex (**B**), or where long-range interactions W^{long} would be mediated by long-ranging horizontal axons, and local interactions W^{local} by the dense vertical/horizontal connection structures within a cortical hypercolumn (**C**). The length scales of input fields are indicated by the size of the image patch sections shown below. Interaction pathways associated with W^{long}, W^{local} and W^{input} are indicated in green, red and blue, respectively. Other links realizing different parts of the model equations (above the schemes) for column u are drawn in black. The putative connection schemes are embedded into sections of primary visual cortex with light and dark gray shading indicating different layers. Note that in our scheme, horizontal interactions originate and terminate in different, but nearby layers as evident from anatomical evidence for layer II-III (McGuire et al., 1991) and layer V-VI (Gilbert and Wiesel, 1983; Hübener et al., 1990), long-ranging axons and that interactions might be indirect by being relayed over intermediary target populations (filled dots) such as inhibitory interneurons.

An alternative picture evolves if we assume that input patches correspond to smaller regions in visual space. Now horizontal interactions within V1 would span over sufficiently long distances to mediate long-range interactions in the model ($W^{long} = C$), while local interactions W^{local} would indeed be local to a cortical (hyper-)column, possibly realized by the dense network linking different cortical layers in a vertical direction (example circuit shown in Fig. 3.8 (C)).

In both discussed scenarios structure and polarity of cortical interactions is compatible with the model: horizontal and feedback connections are orientation-specific, and their effective

interaction can be positive or negative (Hirsch and Gilbert, 1991; Weliky et al., 1995) since they have been found to target both, excitatory and inhibitory neurons (McGuire et al., 1991). It is more difficult, however, to identify the potential locations of populations $\boldsymbol{a}$ and $\boldsymbol{b}$ in the different cortical layers. Two possibilities are shown in Fig. 3.8. The reason why this choice is ambiguous is because indirect input from LGN is provided via layer IV to both superficial and deeper layers (Callaway, 1998), because horizontal axons exists in both layers II-III and layers V-VI (Gilbert and Wiesel, 1983; Rockland and Lund, 1983), and because feedback from higher visual areas also terminates in both superficial and deep layers (Rockland and Pandya, 1979).

Finally, the proposed neural dynamics presents several non-trivial computational aspects, who are essential for producing the contextual effects we obtained. Even though the gradient descent (Eq. (3.10)) and the proposed inference scheme (Eq. (3.22)) have the same fixed points, the latter is much richer in its dynamics, since each reconstruction coefficient is represented by two neural activities who are in addition subject to rectification, and since activities $\boldsymbol{a}$ and $\boldsymbol{b}$ are associated with different neural time constants. In consequence, the effects we describe are most probably caused by a combination of sparse constrained coding and the particular properties of its neural implementation. The fact that all the experimental paradigms we reproduce in our model employ time-varying stimuli makes it hard to disentangle these different factors, since the inference network does never reach a steady state and the largest differences between a 'classic' gradient descent and neural dynamics are expected to show up in those transient epochs.

3.3.5 Contextual effects

Consistently the model reproduces a large variety of contextual phenomena, including size tuning, orientation-contrast effects and luminance-contrast modulations. In particular, all classical and non-classical receptive fields emerge in a fully unsupervised manner by training the model with ensembles of natural images. After training is finished, reproduction of all reported results is possible without change or fine-tuning of parameters, gains or thresholds – just by adhering to the exact visual stimulation procedures as used in the corresponding experimental studies. It is intriguing that also variability of the observed phenomena is reliably reflected in the statistics of model responses. Moreover, when we repeated the contextual-modulation experiments using a more general configuration of the visual field – using four surround patches instead of only one (Appendix. A), we found that using a 'bigger' surround does not affect the agreement between our results and experimental data (the effects at the population-level are reported in Figs. A.2, A.3 and A.4). This close match to experimental findings indicates that the assumed constraints from which dynamics and structure of the model were derived are constructive for providing a comprehensive framework for contextual processing in the visual system.

The nature of the observed effects, being orientation-specific and exhibiting both enhancement and suppression (see Figs. 3.5, 3.6, 3.7), closely mirrors the structures and polarities of local and long-range interactions. Furthermore, they explicitly link functional requirements to the anatomy of the visual system: As already observed in (Zhu and Rozell, 2013), local interactions between similar features are strongly suppressive. They realize competition between alternative explanations of a visual scene which is related to 'explaining away' in Bayesian inference (Lochmann et al., 2012). The effects of long-range interactions depend on the exact stimulus configuration, and on the balance between neural thresholds and the combination of all recurrent inputs in the inference circuit. They serve to integrate features across distances, leading to the enhancement of noisy evidence such as in low-contrast stimuli (Polat et al., 1998),

but also to the suppression of activation by the model finding a simpler explanation for a complex stimulus configuration (i.e., by expressing the presence of multiple collinear line segments in terms of a single contour). This explicit link of natural statistics and cortical dynamics to function is also reflected in psychophysical studies: For example, in natural images an edge co-occurrence statistics being similar to the matrix C was observed and used to quantitatively predict contour detection performance by human subjects via a local grouping rule (Geisler et al., 2001). High-contrast flankers aligned to a low-contrast center stimulus strongly modulated human detection thresholds (Polat and Sagi, 1993), providing facilitation over long spatial and temporal scales of up to 16 seconds (Tanaka and Sagi, 1998). Also detection thresholds of 4-patch stimulus configurations are closely related to natural image statistics (Ernst et al., 2016). In both (Ernst et al., 2016; Polat and Sagi, 1993), the interactions between feature detectors with similar cRF properties are inhibitory for near contexts, and exhibit disinhibitory or even facilitatory effects for far contexts – paralleling the differential effects that local and long-range interactions have in our model.

In parallel to sparse coding, hierarchical predictive coding has emerged as an alternative explanation for contextual phenomena (Rao and Ballard, 1999). The general idea is that every layer in a cortical circuit generates an error signal between a feedback prediction and feedforward inference, which is then propagated downstream in the cortical hierarchy. While being conceptually different on the inference dynamics, the corresponding hierarchical generative model of visual scenes is similar to our paradigm when subjected to spatial constraints.

Besides principled approaches, contextual processing has been investigated with models constructed directly from available physiological and anatomical evidence (Stemmler et al., 1995; Somers et al., 1998; Schwabe et al., 2006). Core circuit of such models is often an excitatory-inhibitory loop with localized excitation and broader inhibition and different thresholds for the excitatory and inhibitory populations, which is similar to our proposed cortical circuits shown in Fig. 3.8 with self-excitation of $\boldsymbol{a}$ and direct excitation on $\boldsymbol{b}$ and broader inhibition provided by W^{local} back onto $\boldsymbol{a}$. Such local circuits are connected by orientation-specific long-range connections, similar to the connections represented by W^{long}, even though they are typically assumed to be more strongly tuned. From these structural similarities we would speculate that contextual effects are caused in both model approaches by similar effective mechanisms.

3.3.6 Outlook

In summary, our paradigm provides a coherent, functional explanation of contextual effects and cortical connection structures from a first-principle perspective, which requires no fine-tuning to achieve a qualitative and quantitative match to a range of experimental findings. For future studies, the model has some important implications:

First, there are experimentally testable predictions. These include the strong dependency of local and long-range interactions on the relative phase of adjacent classical receptive fields. Furthermore, we find two structures emerging in matrix C, namely a diagonal indicating stronger links between neurons with similar orientation preferences, as known from the literature, but also an anti-diagonal indicating enhanced links between neurons with opposite orientation preferences. Since connection probabilities were always reported w.r.t. orientation differences, the latter effect awaits experimental validation. Finally, we expect differences in the statistics of contextual effects between representations $\boldsymbol{a}$ and $\boldsymbol{b}$ to show up when information about the laminar origin of neural recordings is taken into account.

Second, it is formally straightforward to go back from the simplified model with just two

separate input fields to the spatially extended, general scheme and subject it to much 'broader' visual scenes. Moreover, the neural dynamics allows also to address temporal contextual effects, or how neurons would respond to temporally changing contexts in the stimulus such as in 'natural' movies. For example, in simulations we observed strong transient effects shortly after stimulus onset, but a more thorough investigation and comparison to physiological findings is beyond the scope of this paper.

3.4 Methods

3.4.1 Learning and analysis of Φ and C

Variables Φ and C were learned using the procedure outlined in the Results section (Eqs. (3.10)-(3.11)). We sampled input patches of size 16×32 pixels (horizontal configuration) or 32×16 pixels (vertical configuration) from a database of natural images (Olmos and Kingdom, 2004) from which we selected 672 images of size 576×768 pixels in uncompressed TIFF format. Images were first converted from RGB color space to grayscale values and then whitened using a method described in (Olshausen and Field, 1997). The optimization step for $\boldsymbol{a}$ (Eq. (3.10)) was carried out for a batch of 100 image patches with a learning rate of $\eta_a = 0.01$. At the end of each update step for Φ (Eq. (3.11)), the columns of Φ were normalized such that $\|\phi_i\|_2 = 1$. We learned $N = 1024$ feature vectors. Learning was performed with 10^4 iterations each for Φ and C (choosing as learning rates the values $\eta_\Phi = 0.05$ and $\eta_C = 0.01$), after which both dictionary and long-range dependencies matrix were stable. The parameters λ_a and λ_C were set to 0.5 and 0.02. To obtain a better statistics, we repeated learning of the dictionary and of the long-range interactions several times, initializing the simulations with different seeds. The results presented in Figs. 3.4-3.7 are based on $N_{\text{seed}} = 8$ instances of the model.

To parametrize the feature vectors in terms of orientation, spatial frequency, size and location we fitted to each of them a Gabor function of the form

$$g(\theta, \lambda, \sigma_x, \sigma_y, x_0, y_0, \psi) = \kappa \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right) \cos\left(2\pi \frac{y}{\lambda} + \psi\right) + \kappa_0$$

$$\begin{aligned} x &= \quad x_0 \cos(\theta) + y_0 \sin(\theta) \\ y &= -x_0 \sin(\theta) + y_0 \cos(\theta), \end{aligned}$$

where θ is the orientation of the sinusoidal carrier, λ its wavelength, ψ its phase, σ_x and σ_y are the standard deviations of the gaussian envelope, $\kappa > 0$ the contrast and κ_0 an offset. Fitting was done following a standard least square approach.

3.4.2 Simulation of the neural model

The four differential equations that define the neural model (Eqs. (3.22)) were solved numerically with a Runge-Kutta method of order 4 for a time interval of $T = 600$ ms. The time constants τ_h and τ_k were chosen to be 10 ms, close to physiological values of neurons in cortex (Dayan and Abbott, 2001). For analyzing the responses, we discarded the initial transients and averaged over single cell activities over the last 333 ms, a period of time that allowed a complete cycle of the stimulus drifting with a temporal frequency of 3 Hz, being the average preferred speed for cortical neurons (Foster et al., 1985).

To ensure positivity of neural responses, in addition to the differential equations Eqs. (3.22) we had to introduce a linear threshold operation (Eqs. (3.13) and (3.19)). In contrast, no constraint is imposed on the sign of $\boldsymbol{a}$ and $\boldsymbol{b}$ in the generative model (Eqs. (3.4)-(3.7)), nor by

the optimization equation (3.10). To make the neural model consistent with the generative model, we therefore duplicated the number of neurons by introducing ON- and OFF units (see Subsection 3.2.2). In addition, we considered for all dictionary elements ϕ_i also their mirrored versions $-\phi_i$ and we split the long-range interactions into positive and negative contributions $C^+ = \max(C, 0)$ and $C^- = \min(C, 0)$ via

$$\Phi \leftarrow \begin{bmatrix} \Phi & -\Phi \end{bmatrix} \in \mathbb{R}^{M \times 2N} \text{ and} \tag{3.26}$$

$$C \leftarrow \begin{bmatrix} C^+ & C^- \\ C^- & C^+ \end{bmatrix} \in \mathbb{R}^{2N \times 2N}. \tag{3.27}$$

For selecting cells well-tuned and well-responding to stimuli centered in one input patch (see Subsection 3.2.4), all units were first stimulated with a set of small drifting sinusoidal gratings centered at $\boldsymbol{r}^u$ with $r_c = 2$ pixels and $k_c = 1$. We varied θ_c from 0 to π in steps of π/N_θ ($N_\theta = 36$) and the spatial frequency f_c from 0.05 to 0.35 cycles/pixel in steps of 0.025. We then selected for each neuron the preferred orientation and preferred spatial frequency. A unit was said to be responsive if its peak response was at least 10% of the maximum recorded activity. We determined orientation selectivity by computing, for each unit n, the complex vector average

$$z_n = \sum_{k=0}^{N_\theta - 1} a_n(\theta_k) e^{2i\theta_k} \Big/ \sum_{k=0}^{N_\theta - 1} a_n(\theta_k), \text{ for } \theta_k = \frac{2\pi k}{N_\theta},$$

and we considered tuned those neurons for which it was $|z_n| > 0.85$, corresponding to a tuning width of approximately 20 degrees half-width. With these selection criteria, we were left with 490 cells from all N_{seed} instantiations of the model.

3.4.3 Selection of orientation contrast tuning classes

When we quantified the effect of cross-orientation stimulation, we pooled responses of units exhibiting the same qualitative behavior (Fig. 3.6 (C, D)). To determine which behavior a unit showed we first computed, for each unit n with preferred orientation $\theta^\star$ degree, the average response to the compound stimulus when the surround orientation was close to $\theta^\star$

$$\bar{a}_n^\star = \frac{1}{10} \int_{\theta^\star - 5^\circ}^{\theta^\star + 5^\circ} a_n(\theta_a) d\theta_a$$

and when the surround orientation was near-oblique

$$\bar{a}_n = \frac{1}{10} \int_{\theta^\star - 20^\circ}^{\theta^\star - 10^\circ} a_n(\theta_a) d\theta_a + \frac{1}{10} \int_{\theta^\star + 10^\circ}^{\theta^\star + 20^\circ} a_n(\theta_a) d\theta_a.$$

The unit was considered to show iso-orientation suppression if $\bar{a}_n - \bar{a}_n^\star > \varepsilon$, release from suppression if $\bar{a}_n^\star - \bar{a}_n > \varepsilon$ and untuned suppression in all other cases ($\varepsilon = 0.05$).

3.4.4 Constants and parameters

Parameters used in numerical simulations are summarized in Table 3.1.

The code to implement the model is available at the following link
`https://github.com/FedericaCapparelli/ConstrainedInferenceSparseCoding`.

Table 3.1: Parameters values used in simulations.

Size tuning			
θ_c	preferred	θ_a	-
ω_c	preferred	ω_a	-
r_c	from 2 to 32 in steps of 1	r_a	-
k_c	1	k_a	-

Orientation-contrast (center-only)			
θ_c	from 0 to π in steps of $\pi/36$	θ_a	-
ω_c	preferred	ω_a	-
r_c	optimal	r_a	-
k_c	1	k_a	-

Orientation-contrast (center-surround)			
θ_c	preferred	θ_a	from 0 to π in steps of $\pi/36$
ω_c	preferred	ω_a	preferred
r_c	optimal	r_a	∞
k_c	1	k_a	1

Luminance-contrast (center-only)			
θ_c	preferred	θ_a	-
ω_c	preferred	ω_a	-
r_c	optimal	r_a	-
k_c	from 0.1 to 1 in steps of 0.1	k_a	-

Luminance-contrast (center-surround)			
θ_c	preferred	θ_a	preferred
ω_c	preferred	ω_a	preferred
r_c	optimal	r_a	∞
k_c	from 0.1 to 1 in steps of 0.1	k_a	1

3.4.5 Acknowledgement

This chapter including figures was published in similar form as

F. Capparelli, K. R. Pawelzik, and U. A. Ernst. “Constrained inference in sparse coding reproduces contextual effects and predicts laminar neural dynamics". PLoS computational biology 15.10 (2019): e1007370.

The text underwent minor changes for consistency with other chapters.

4 A model of spontaneous activity

4.1 Introduction

Experiments conducted across a variety of species, with different methods and different experimental setups revealed a striking resemblance between patterns of activity occurring spontaneously and under sensory stimulation in primary visual cortex. The emergence of spatially inhomogeneous patterns of spontaneous activity on the scale of several hundred micrometers has been long known (Arieli et al., 1995) and subsequent experiments helped characterizing their appearance and their dynamics. For instance, using voltage sensitive dyes in the visual cortex of anesthetized cats and monkeys, Grinvald and colleagues reported maps of spontaneous activity showing a high degree of correlation with orientation maps evoked by oriented stimuli commonly used to investigate early visual processing (Kenet et al., 2003; O'hashi et al., 2017; Omer et al., 2018). Recordings performed in primary visual cortex of *awake* ferrets and monkeys (Smith et al., 2018; Omer et al., 2018) also reported spontaneous widespread modular correlation patterns, tightly related to the functional organization of iso-orientation domains. When spontaneous events occur, such iso-orientation domains become active either simultaneously or in a sequence (Smith et al., 2018; Omer et al., 2018), lasting several hundred of milliseconds (Kenet et al., 2003; O'hashi et al., 2017; Smith et al., 2018; Omer et al., 2018). In the orientation dimension, the transitions from a state to the next can be either smooth, when a state with a certain orientation is followed by a state with a proximal orientation, or discontinuous (O'hashi et al., 2017).

The mechanisms responsible for generating these spontaneous events are still not clearly understood. Even though a causal role for retinal and thalamic feedforward inputs in establishing correlated modular structures cannot be ruled out, experimental evidence suggests that such complex states are shaped and expressed through intrinsic cortical mechanisms. Theoretical investigations have corroborated the idea that lateral orientation-specific connections have a great relevance to the large-scale spatial organization of both sensory representation in primary sensory cortices and spontaneous activity (Ernst et al., 2001; Goldberg et al., 2004; Blumenfeld et al., 2006). Two models in particular have focused on explaining the similarities between spontaneous end evoked states (Goldberg et al., 2004; Blumenfeld et al., 2006). Both models propose a simplified V1 network where neural populations are arranged according to a ring architecture and are linked by a cosine-shaped connectivity pattern, connecting excitatorily distant neurons with similar orientation preferences. Linear stability analysis performed on the dynamics on either one of the models reveals the existence of two scenarios, depending on how parameters are tuned (cortical gain and input strength), one where an homogeneous fixed point is stable and spontaneous states represent fluctuations around a single background cortical state and another where the cortex wanders among multiple attractor states, namely cortical maps which encode various visual attributes. Both models predict that in the multiple-attractors regime, the network will exhibit a random walk on the manifold of attracting states. However, the dynamics observed in experiments also exhibits abrupt transitions in orientation space,

usually to the orthogonal state. To explain the emerging and decay of spontaneous states a more comprehensive modeling approach is necessary, where the role of noise might be crucial. Another aspect that was not thoroughly explored is the presence of *mixed* states, that is patterns that are composed of states tuned to different orientations in different cortical regions, or localized states, whose lateral spread is less than the whole imaged area. Mosaic states were found occasionally in recordings under anesthesia (O'hashi et al., 2017), and could be obtained thanks to symmetry breaking in the coupling matrix (e.g. introducing anisotropies (Blumenfeld et al., 2006)). Localized states spanning a few hypercolumns match the description of maps obtained with awake animals (Omer et al., 2018; Smith et al., 2018), where widespread activity is more rare, and their spatial extent could be mediated by long-range horizontal axons. Mosaic and localized states are more compatible with the activity patterns observed in non-anesthetized animals, therefore it is interesting to understand which interactions determine them, especially if we wish to exploit ongoing cortical states to insert artificial signals.

In this Chapter, we introduce a structurally simple model with local and long-range interactions capable to exhibit pattern formation. Given the analytical tractability of the model, using linear stability analysis, we first study pattern formation analytically, identifying conditions and parameters under which biophysically realistic patterns arise. Next, we introduce a stochastic dynamics to allow spontaneous pattern formation and decay, and we perform a numerical investigation of the network for different noise levels. Finally, we quantify several properties that are important for comparing the model to experimental data, such as the emergence and decay probabilities, average duration of states, their lateral spread in cortical space and the prevalence of mixed states.

The analytical and numerical understanding gained on the role of parameters in generating realistic spontaneous spatiotemporal patterns of activity will be used in Chapter 5 for developing a stimulation paradigm that might be employed in a cortical prosthetic device to evoke orientation-specific percepts in a blind subject.

4.2 Results

To investigate the dynamics of spontaneous activity in visual cortex and its underlying mechanisms, we start by building a minimalistic model that is able to generate the complex, highly structured spatial patterns described in Section 2.6 while at the same time retain some analytical tractability. In the course of this chapter, when it will become necessary to increase the complexity of the model to account for a larger number of experimentally observed facts, we will progressively modify the model, including more biologically plausible features.

4.2.1 Neuronal populations: dynamics and interactions

The elementary computational unit that we consider is a population of neurons sharing the same orientation preference. We denote its activity at time t by $A(r,t)$, where r is the position that the population occupies in the visual cortical domain we are analyzing. For practical purposes, we consider the following discretization of the real axis in intervals of length 2π

$$(-\infty,\infty)=\bigcup_{j\in\mathbb{Z}}[(2j-1)\pi,(2j+1)\pi), \tag{4.1}$$

so we can identify each hypercolumn (indexed by j) with a fundamental period and can parametrize each population with its orientation preference $\theta=\frac{r}{2} \mod \pi$.

The activity of each population is described by a Wilson-Cowan type equation (Wilson and Cowan, 1973)

$$\tau \frac{\partial A}{\partial t}(r,t) = -A(r,t) + g\Big([W \star A](r,t) + I^{\text{ext}}(r,t)\Big) \tag{4.2}$$

where $W \star A$ and I^{ext} represent the recurrent and external input, $\star$ denotes the convolution operation

$$[W \star A](r,t) = \int_{-\infty}^{\infty} W(r-r')A(r',t)dr' \tag{4.3}$$

and g is a linear-threshold gain function

$$g(y) = [\gamma(y-b)]^+, \quad \gamma, b > 0. \tag{4.4}$$

The coupling matrix $W = W(\Delta r)$ that defines the recurrent interactions between two populations of neurons at distance $\Delta r = |r - r'|$ is given be the sum of two terms, operating on two different spatial scales, termed 'local' and 'long', that are given by

$$W^{\text{local}}(\Delta r) = -J_I \frac{1}{\sqrt{2\pi}\sigma_I} \exp\Big(-\frac{|r-r'|^2}{2\sigma_I^2}\Big) \tag{4.5}$$

$$W^{\text{long}}(\Delta r) = J_E \frac{1-\Lambda}{1+\Lambda} \frac{1}{\sqrt{2\pi}\sigma_E} \sum_{j=-\infty}^{\infty} \exp\Big(-\frac{|r-r'-2j\pi|^2}{2\sigma_E^2}\Big)\Lambda^{|j|}. \tag{4.6}$$

While the coupling W^{local} is inhibitory and acts only locally, W^{long} links similarly tuned populations even if they are distant from each other, mimicking the long excitatory horizontal axons in V1 (Gilbert and Wiesel, 1989; Bosking et al., 1997; Malach et al., 1993). Both terms are modelled as Gaussians with length scales σ_I and σ_E, for which we assume $\sigma_I > \sigma_E$. The excitatory coupling is defined as a sum of Gaussians placed periodically on the real axis, scaled by Λ, a parameter that controls how fast the interaction strength decays as a function of distance from the presynaptic population, with $0 \leq \Lambda < 1$. The constants J_I and J_E represent the total coupling strength (note that Eqs. (4.5) and (4.6) are normalized such that their integrals are $-J_I$ and J_E respectively). A sketch of the recurrent interactions is given in Fig. 4.1.

Note that, by setting $\Lambda = 0$, the above defined lateral interactions reduce to a Mexican-Hat shape, and the model becomes identical to the one used in (Ernst et al., 2001).

4.2.2 Linear stability analysis

When subject to a constant supra-threshold input $I^{\text{ext}}(r,t) = I > b$, the differential equation (4.2) has a non-negative homogeneous fixed point

$$A^\star = \frac{\gamma(I-b)}{1+\gamma J_I - \gamma J_E} \tag{4.7}$$

if

$$J_I > J_E - 1/\gamma. \tag{4.8}$$

To investigate its stability, we consider a small perturbation δA around the fixed point, and we study the behavior of Equation (4.2) for $A(r,t) = A_0 + \delta A(r,t)$, that is

$$\tau \frac{\partial \delta A}{\partial t}(r,t) = -\delta A(r,t) + \gamma\big((W^I \star \delta A)(r,t) + (W^E \star \delta A)(r,t)\big). \tag{4.9}$$

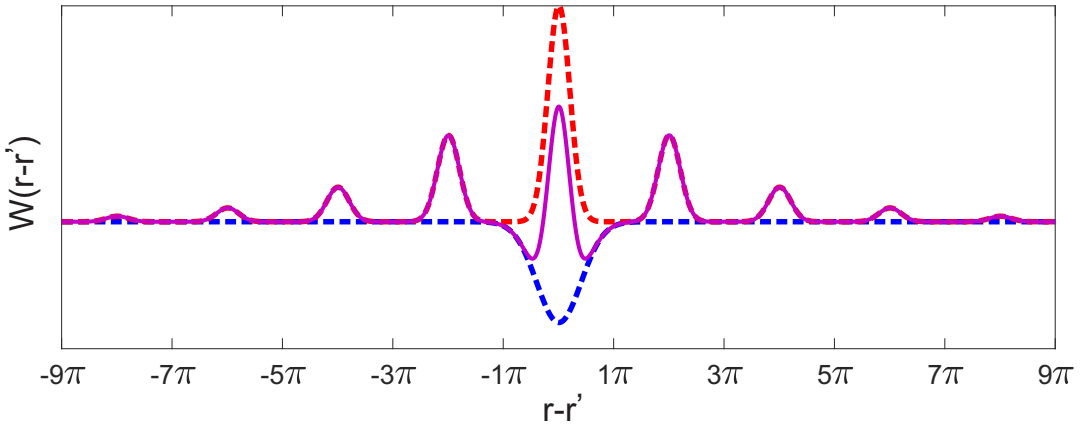


Fig. 4.1: Recurrent interactions. Coupling strength as a function of the distance between the populations. Excitatory and inhibitory interactions are drawn in red and blue, the net interaction in purple.

This analysis becomes easier in the spatial-frequency domain, since taking the Fourier transform of Eq. (4.9) results in a system of decoupled ordinary differential equations

$$\tau \frac{\partial \delta \hat{A}}{\partial t}(\omega, t) = \nu_\Lambda(\omega) \delta \hat{A}(\omega, t) \tag{4.10}$$

where $\nu_\Lambda(\omega)$ are the eigenvalues of the linearized dynamics

$$\nu_\Lambda(\omega) = -1 - \gamma J_I \exp\Big(-\frac{\sigma_I^2 \omega^2}{2}\Big) + \gamma J_E \frac{1-\Lambda}{1+\Lambda} \sum_{j=-\infty}^{\infty} \Lambda^{|j|} \exp\Big(-\frac{\sigma_E^2 \omega^2}{2} + (2\pi i)\omega j\Big). \tag{4.11}$$

The sign of $\Re(\nu_\Lambda(\omega))$ for a particular frequency ω (also called 'mode'), determines whether a perturbation with that spatial frequency will decay ($\Re(\nu_\Lambda(\omega)) < 0$) or grow ($\Re(\nu_\Lambda(\omega)) > 0$) exponentially. For this reason, studying the sign of $\Re(\nu_\Lambda)$ gives information about the stability of the fixed point (Strogatz, 2018). In particular:

(i) If $\Re(\nu_\Lambda(\omega)) < 0$ for all ω, the fixed point is stable;
(ii) If there is at least one ω for which $\Re(\nu_\Lambda(\omega)) > 0$, then the fixed point is unstable.

The infinite sum and the oscillatory component given by the complex exponential contained in Equation (4.11) make $\Re(\nu_\Lambda(\omega))$ difficult to treat analytically (a representation of the function is given in Fig. 4.2). However, it can be shown that Eq. (4.11) admits the following closed-form expression

$$\Re(\nu_\Lambda(\omega)) = -1 - \gamma J_I \exp\Big(-\frac{\sigma_I^2 \omega^2}{2}\Big) + \gamma J_E \exp\Big(-\frac{\sigma_E^2 \omega^2}{2}\Big) \frac{(\Lambda-1)^2}{\Lambda^2 - 2\Lambda\cos(2\pi\omega) + 1} \tag{4.12}$$

and that the envelope of such family of functions is given by the two extreme cases $\Lambda = 0$ and $\Lambda = 1$ (see Appendix B). In particular, for all $\omega > 0$, we have that

$$\Re(\nu_1(\omega)) \leq \Re(\nu_\Lambda(\omega)) \leq \Re(\nu_0(\omega)) \tag{4.13}$$

for all $\Lambda \in (0, 1)$.

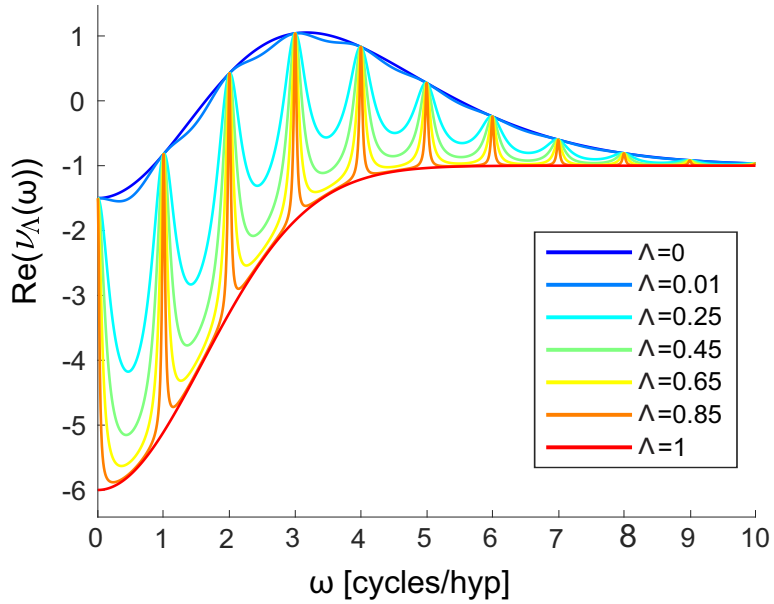


Fig. 4.2: Real part of the eigenvalues spectrum. Different colors correspond to different values of Λ. All curves are contained in the envelope formed by $\Lambda = 0$ and $\Lambda = 1$. Local maxima occur at integer values of ω.

The second inequality in Eq. (4.13), together with condition (ii), implies that the stability of the fixed point for the case $\Lambda = 0$ is a sufficient condition for the stability for $\Lambda > 0$. We will therefore limit ourselves to find conditions under which (ii) holds when $\Lambda = 0$. Let's call

$$S = \frac{\sigma_I^2}{\sigma_E^2} > 1$$

the ratio of the inhibitory and excitatory length scale. Assuming that the fixed point exists, studying the sign of the first derivative of the function

$$f(\omega) := \Re(\nu_0(\omega)) = -1 - \gamma J_I \exp\Big(-\frac{\sigma_I^2\omega^2}{2}\Big) + \gamma J_E \exp\Big(-\frac{\sigma_E^2\omega^2}{2}\Big) \tag{4.14}$$

for $\omega > 0$, yields two possibilities (more extensive computations are given in Appendix B):

- When $J_I < J_E S^{-1}$, $f(\omega)$ assumes its maximum value at $\omega = 0$. Since such maximum is negative, the fixed point is stable.
- When $J_I > J_E S^{-1}$, $f(\omega)$, has ones local minimum at $\omega = 0$ and an absolute maximum at $\omega = \omega^\star = \sqrt{\log\Big(\frac{J_I\sigma_I^2}{J_E\sigma_E^2}\Big)\Big(\frac{\sigma_I^2-\sigma_E^2}{2}\Big)^{-1}}$. Therefore the fixed point is stable if and only if $f(\omega^\star) < 0$, which is true if and only if $J_I > (J_E)^S(S-1)^{(S-1)}S^{-S}$.

The conditions stated above divide the phase space in three regions (Fig. 4.3 (A)), associated with three dynamical regimes. In the lower region the fixed point does not exist and the network activity grows without bound (divergent phase). This is the case when the mode $\omega = 0$ is unstable, i.e. $\Re(\nu_\Lambda(0)) > 0$. In the upper region, where the local inhibition is sufficiently strong, the homogeneous fixed point is stable and every perturbation decays exponentially (linear phase). In this regime, the steady-state activity profile looks flat. If the long-range excitation becomes strong enough, any perturbation of an otherwise constant external input

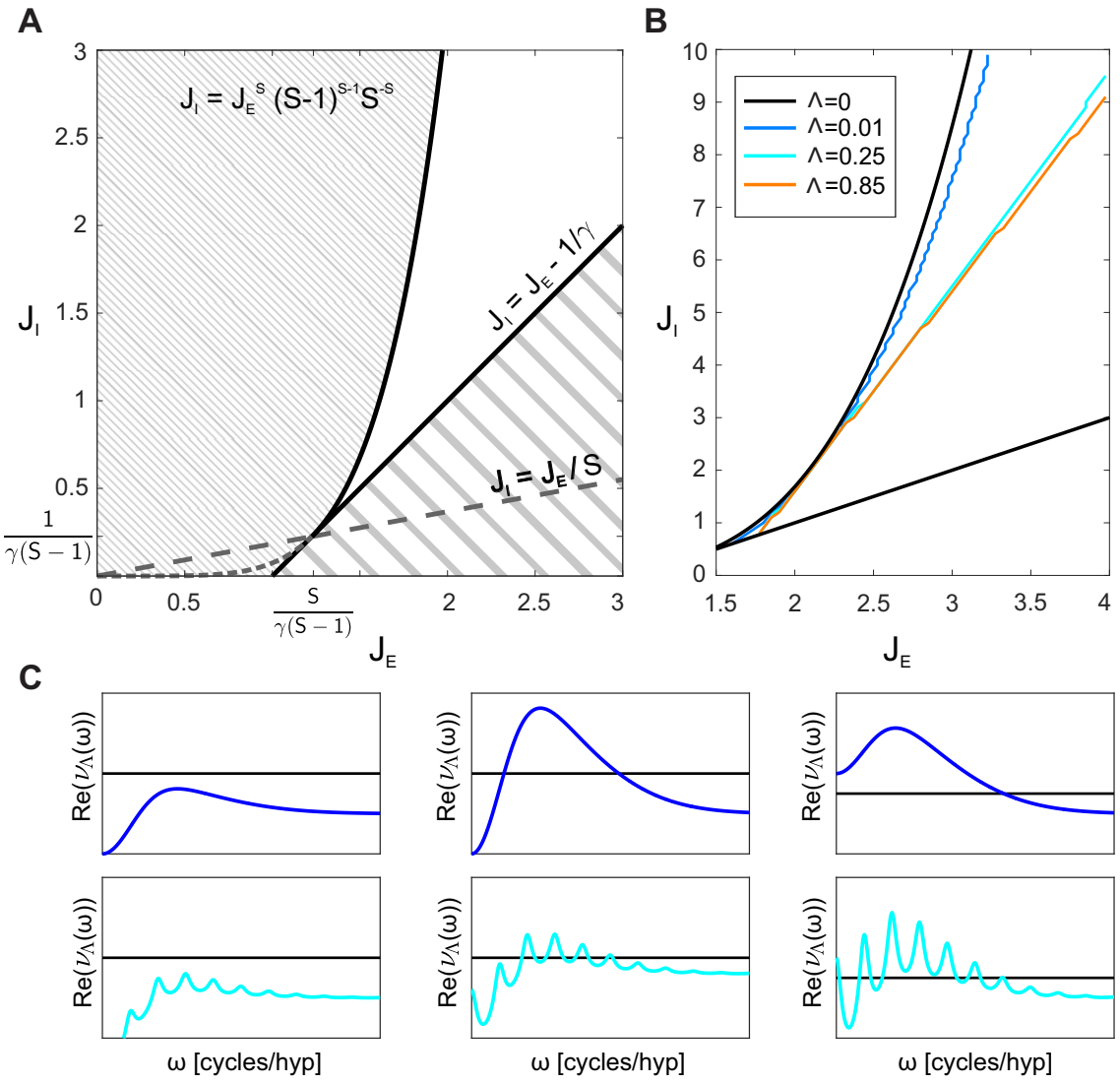


Fig. 4.3: (**A**) The black curves are the boundaries between different qualitative regimes in the phase space spanned by the excitatory and inhibitory coupling strength (case $\Lambda = 0$), termed *linear* (densely shades area), *marginal* and *divergent* (coarsely shaded area) regime. (**B**) When $\Lambda \neq 0$, the boundary between linear and marginal phase deviates from the analytical expression defined by Eq. (4.15) (drawn in black). (**C**) Qualitative shape of the eigenvalues in the linear (left), marginal (middle) and divergent (right) regime for $\Lambda = 0$ (top) and $\Lambda > 0$ (bottom).

leads to a regime where the system converges into a spatial pattern where localized populations of neurons are active. In this phase, either the spatial pattern stabilizes itself (marginal phase) or cannot stabilize and grows without bounds. These three regions are associated with a particular qualitative shape of the eigenvalues, as exemplified in Fig. 4.3 (C).

In the following sections we will consider a parameter range where the inhibition strength is sufficiently high, namely $J_I > 1/(\gamma(S-1))$, so that the boundaries between linear and marginal

phase and between marginal and divergent phase are given, respectively, by

$$J_I = (S-1)^{S-1}\left(\frac{J_E}{S}\right)^S \tag{4.15}$$

and

$$J_I = J_E - \frac{1}{\gamma}.$$

If the length scale of long-range interactions Λ is increased above zero, the curve defined by Eq. (4.15) is only an upper bound of the true boundary between linear and marginal phase. One can indeed notice from Fig. 4.2 that there are modes ω for which $\Re(\nu_0(\omega))$ is positive, but $\Re(\nu_\Lambda(\omega))$ isn't. The deviations between the boundaries can be quantified numerically and are shown for different values of Λ in Fig. 4.3 (B).

4.2.3 Network state in the marginal phase

For illustrative purposes, it is worth showing an example of how neural activity develops over time in the marginal phase. To perform simulations of the presented dynamical system, we need to go away from a mathematical formulation in terms of continuous variables and drop the hypothesis of an infinitely extended system. So, we model the cortical space $\mathscr{C}$ with periodic boundary conditions, by considering the finite length interval $\mathscr{C} = (0, 2N_H\pi)$. We interpret $\mathscr{C}$ as composed by N_H fundamental periods of length 2π, each corresponding to an hypercolumn and containing N_P neural populations. We discretized this space with spatial resolution $\Delta x = \frac{2\pi}{N_P}$, which allows to denote the activity of a population at position $x_n = n\Delta x$, $n = 1, 2, \ldots, N_H \cdot N_P$ with A_n.

Fig. 4.4 (A) shows five successive snapshots of activity for a single hypercolumn. The activity organizes from a random initial condition (uppermost panel) into a spatial profile with three localized bumps per hypercolumn (mid panels); as time grows, the amplitude of one of those bumps decreases until it disappears completely and the network settles into a pattern with only two bumps per hypercolumn (lowermost panel). The spatio-temporal profile for the whole system ($N_H = 10$ hypercolumns) is reported in Fig. 4.4 (B). As we explain in the next section, the number of bumps – or, better, the number of cycles – emerging in the network can be predicted in advance from the spectrum of eigenvalues (Fig. 4.4 (C)). Two modes, in particular, are important to determine this number: the frequency associated with the maximal amplitude that, in this example, corresponds to 3 and the frequency associated to the minimal non-zero amplitude, close to 2.

4.2.4 Plausible parameter range

To tune the model into a regime in which the network dynamics is comparable with V1 ongoing activity, we need to focus on the dynamical regime where pattern formation happens. In particular, to explain the similarity between spontaneous activity patterns and single orientation maps, we need to restrict the parameter to a range that ensures that the emergent number of cycles per hypercolumn is one.

To do this, we once again set $\Lambda = 0$ and we consider the insights that we can gain from the linear stability analysis. Typically (Scherf et al., 1999; Wolf et al., 2000), marginally stable dynamical systems initially develop an activity pattern dominated by frequencies with high amplitude and then settle into patterns with lower frequencies. As we observed in Fig. 4.4, the expected number of cycles can be deduced from the mode at which the spectrum of eigenvalues

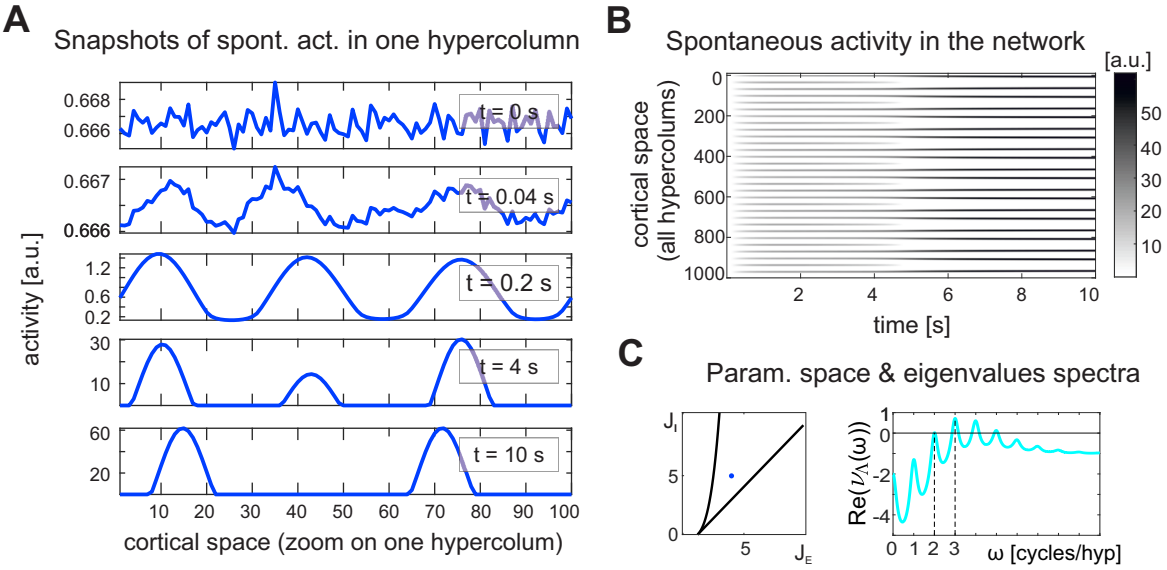


Fig. 4.4: Example of pattern formation. (**A**) Activity pattern emerging in one hypercolumn at different points in time. (**B**) Activity of the entire network over time. In this simulation we set $N_H = 10$, $N_P = 100$, $J_I = 5$, $J_E = 4$, $\sigma_E = 0.05$, $\sigma_I = 0.1$, $\Lambda = 0.25$, $I^{\text{ext}} = 10$.

$f(\omega)$ (Eq. (4.14)) is maximized and from the mode at which it becomes positive. We will denote these two variables, respectively, with $\omega_{\max}$ and $\omega_{\min}$. Their values is determined by four parameters, namely J_E, J_I, σ_E and J_I, and is shown in Fig. 4.5 (first and last column) in the parameters space spanned by excitatory and inhibitory coupling strength for three different values of the ratio S.

The actual number of cycles n_{cycles} can be estimated numerically by simulating the differential equation (4.2) until a time $t = T$ and then counting the number of cycles per hypercolumn that the network shows. Figure 4.5 shows the value of n_{cycles} computed for three values of T (mid columns) in a network with Mexican-hat connectivity ($\Lambda = 0$, Fig. 4.5 (A-C)) and in a network with long-range connectivity ($\Lambda > 0$, Fig. 4.5 (D)). For the values of spatial scales chosen in this simulation, the number of cycles lies in the same range as $\omega_{\max}$ and $\omega_{\min}$, varying from 0 to 5 across the space spanned by J_E and J_I, but shows a temporal evolution. In fact, comparing such numerical estimates with the analytical values shows that, for small values of T, n_{cycles} initially assumes values close to $\omega_{\max}$; as we let the dynamics evolve for longer periods, n_{cycles} progressively changes and gets closer to $\omega_{\min}$.

With respect to the parameters σ_E and σ_I, the above observations do not vary. Increasing the ratio between the range of excitation and inhibition has the only effect of moving the boundary curve defined by Eq. (4.15), effectively expanding the region where the system exhibits pattern formation.

Changing the value of Λ from zero to non-negative values also preserves the same relation between n_{cycles}, $\omega_{\max}$ and $\omega_{\min}$. However, setting $\Lambda > 0$ causes n_{cycles} to change less smoothly as J_E is increased and to assume only discrete values. This is particularly evident when considering a cross-section of the parameter space (Fig. 4.5, black and blue curves represent the number of bumps per hypercolumn computed in a network with Mexican-hat connectivity (black) and with long-range connectivity (blue)).

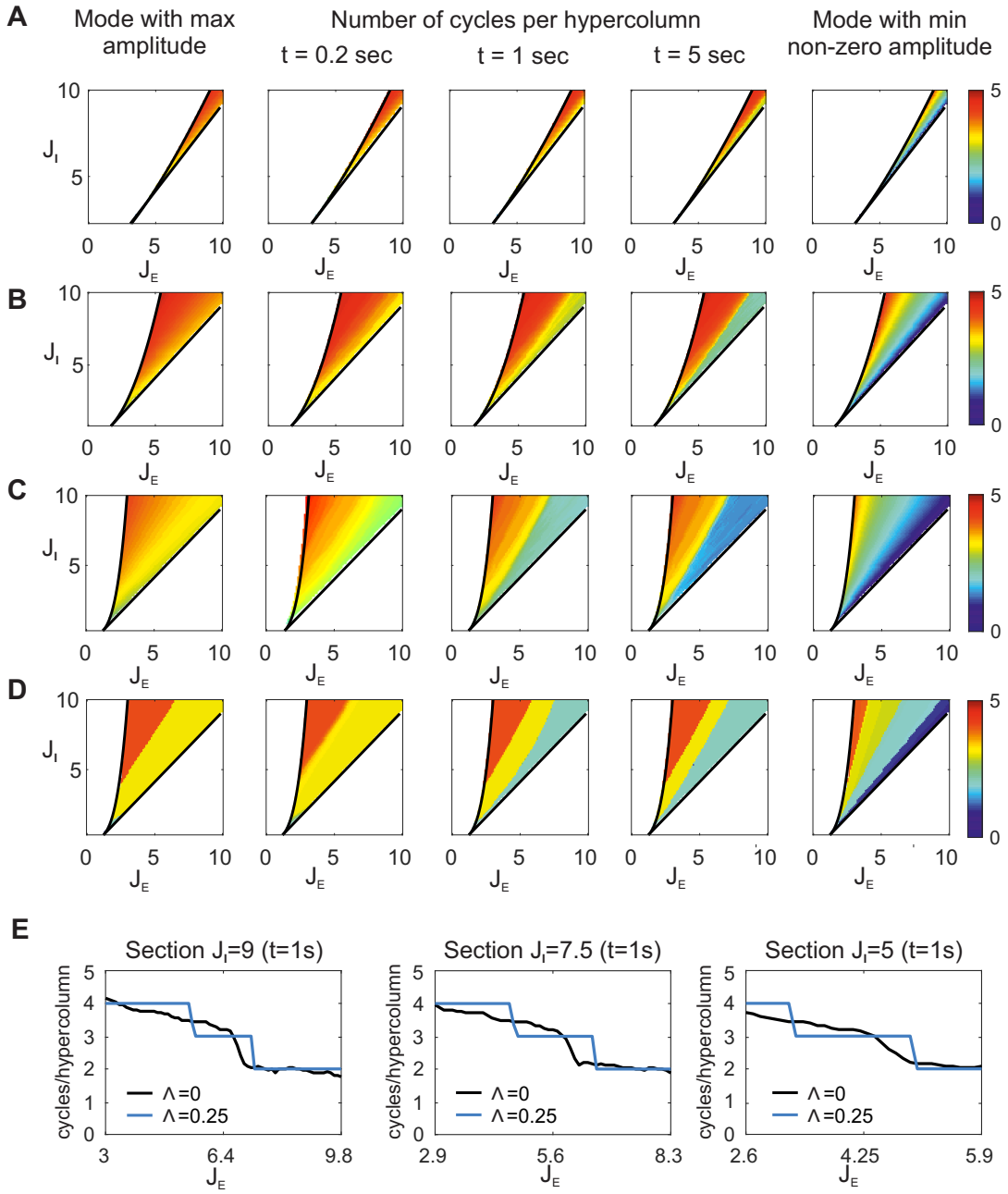


Fig. 4.5: (**A-D**) From left to right, mode with the highest amplitude, number of cycles per hypercolumn computed at $t = 0.2, 1$ and 5 s, mode with the lowest amplitude with (**A**) $\sigma_E = 0.05$, $\sigma_I = 0.06$, $\Lambda = 0$, (**B**) $\sigma_E = 0.05$, $\sigma_I = 0.07$, $\Lambda = 0$, (**C**) $\sigma_E = 0.05$, $\sigma_I = 0.1$, $\Lambda = 0$, (**D**) $\sigma_E = 0.05$, $\sigma_I = 0.1$, $\Lambda = 0.25$. (**E**) Cross-sections of the images shown in (**C**) and (**D**).

To find parameter combinations that produce activity patterns with one cycle per hypercolumn only, we need to consider a simulation time close to the time scale of the experiments we wish to reproduce. Since the average durations of spontaneous events is reported to be approximately 200 ms (O'hashi et al., 2017; Omer et al., 2018), it makes sense to look for parameters for which $\omega_{\max} \approx 1$. This can be achieved by finding the right spatial scales and in particular increasing the value of σ_I. This is reasonable considering that σ_I represents the length scale of local inhibition: if it's large enough, it is not possible that two bumps of activity emerge in a

single hypercolumn. A plausible choice is obtained by setting $\sigma_E = 0.15$ and $\sigma_I = 0.35$. Such values will be used almost exclusively in subsequent simulations.

4.2.5 Dynamics of spontaneous states

The model described so far produces static patterns in the limit $t \to \infty$. Experimental observations, however, present spontaneous events that are localized in space, have a finite durations and exhibit a characteristic state-switching dynamics (Kenet et al., 2003; O'hashi et al., 2017; Smith et al., 2018; Omer et al., 2018). To reproduce these findings we need to introduce in the system a noise source capable of triggering the emergence of orientation-tuned states, morphing them in orientation space and also disrupting them.

There are many ways in which a stochastic version of the model described by Eq. (4.2) can be realized. A straightforward way consists in interpreting the time course of the variable A as the underlying time-dependent rate of an inhomogeneous Poisson process. In particular, the spikes produced at time t by a population of neurons at position r are generated by a time-varying poisson process with rate $A(r,t)$. These spikes are then used to compute the recurrent input to other populations in the network. The new stochastic dynamics is given by

$$\begin{cases} \tau \frac{\partial A}{\partial t}(r,t) = -A(r,t) + g\big([W \star \rho](r,t) + I^{\text{ext}}(r,t)\big) \\ \rho(r,t) \sim \text{Poisson}(A(r,t)). \end{cases} \tag{4.16}$$

A first interpretation of this dynamics is that it implements noise at the level of synaptic transmission. A second and equally plausible interpretation is that A constitutes an internal average activation variable of a cortical population, and that spikes are generated with an average rate proportional to that activation. Here the stochasticity lies in the fact that we do not know which of the (equally activated) neurons lies currently closer or more distant to the firing threshold.

Another way of introducing stochasticity into Eq. (4.2) is to use a temporally decaying Gaussian input. The new stochastic dynamics, in this case, is given by

$$\begin{cases} \tau \frac{\partial A}{\partial t}(r,t) = -A(r,t) + g\big([W \star A](r,t) + I^{\text{ext}}(r,t)\big) \\ \frac{dI^{\text{ext}}}{dt} = -\frac{I^{\text{ext}}}{\tau'} + \frac{\mu}{\tau'} + \frac{\sigma}{\sqrt{\tau'}}\eta, \end{cases} \tag{4.17}$$

where μ and σ are the drift and the diffusion coefficient of the external input, τ' is its time constant and η follows a standard Gaussian distribution.

While in the model defined by Eq. (4.17) the noise is caused by the input, in the model defined by Eq. (4.16) the noise intrinsically depends on the mean activation level $A^\star$ (Eq. (4.7)). We will thus call these two kinds of noise *input noise* and *intrinsic noise*. Input noise can be quantified as the coefficient of variation,

$$C_v = \frac{\sigma}{\mu}$$

and can be easily controlled by varying σ. Since intrinsic noise is generated as a Poisson process, the variance of the activity is equal to $A^\star$. We can thus use $A^\star$ to regulate the noise, by choosing the appropriate I^{ext}.

An example of spontaneous patterns generated by the Equations (4.17) is given in Fig. 4.6 (A), where we depicted the orientation that emerges over time in each stimulated hypercolumn, color coded as indicated in the colorbar (a similar picture emerges using the dynamics given by Equations (4.16)). As in the deterministic model, the dynamics is able to produce localized

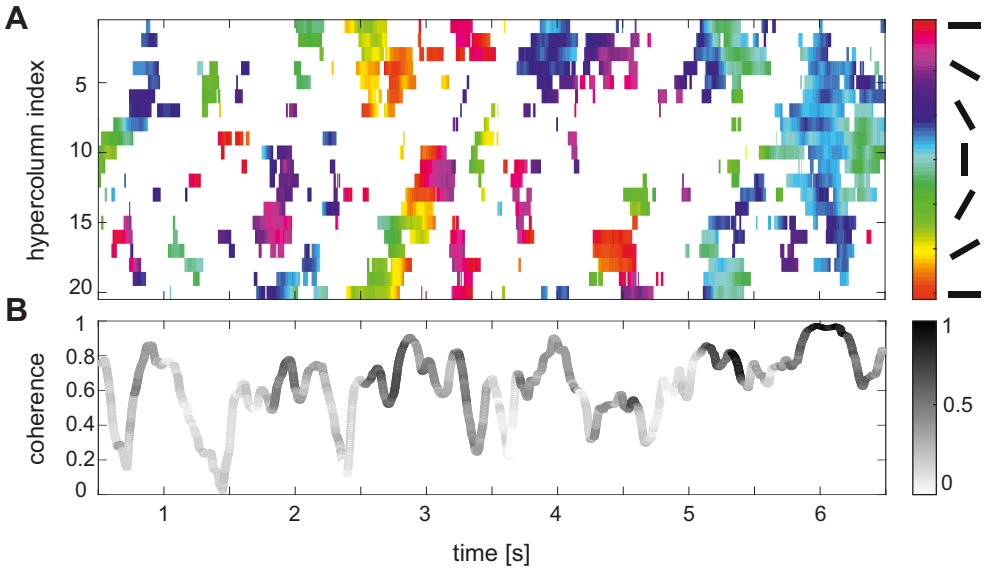


Fig. 4.6: Stochastic dynamics. (**A**) Example of spontaneously emerging and decaying oriented states. Tuning of the state is indicated by the corresponding color, white patches indicate times and hypercolumns at which there was no state detected. The procedure with which the presence of cortical states is assessed is described in detail in the Methods section. Briefly, a state is detected when the absolute value of the complex average of the activity $A(r,t)$ in each hypercolumn surpasses a threshold. (**B**) Coherence of the states shown in (**A**). Here, the contrast of the curve is scaled with the number of hypercolumns that form a state, with darker shades corresponding to states spread over bigger cortical regions.

activity-bumps in the network, resembling the spontaneous state observed in experiments (Kenet et al., 2003). The variability in the noisy input, however, breaks the periodicity of the bumps, allowing the development of a rich collection of phenomena, similar to those reported for anesthetized and awake animals (O'hashi et al., 2017; Smith et al., 2018; Omer et al., 2018). For example, over time, the peak of activity can shift both in cortical and in orientation space, creating a dynamic emergence and decay of cortical states (see highlighted areas in Fig. 4.6 (A)). Typically, those states span a few hypercolumns and occasionally extend over the whole cortical space. In addition, in different regions of the simulated cortex, populations with different orientation preferences might prevail, creating a 'mosaic' of oriented states. We quantified the prevalence of these mixed states with a index that we call 'coherence', whose values range from 0 to 1. High values of coherence indicate that the orientations to which single hypercolumns are tuned are similar (coherent) throughout the cortex, corresponding to a global cortical state. Low values of coherence indicate that the orientations to which single hypercolumns are tuned are different from each other, corresponding to a mosaic of local states.

4.2.6 Properties of spontaneous states

To investigate the possible mechanisms through which spontaneous oriented states emerge and to analyze how the properties of such states depend on the parameters of the model, we simulated the models given by Equations (4.16) and (4.17) for different levels of noise.

Intuitively, at low levels of noise, the stochastic dynamics will behave similarly to its deterministic

counterpart. With higher level of noise, however, spatial-frequency modes can be amplified, causing the formation of activity patterns over larger regions of parameter space than in the deterministic case.

To quantify this idea, we computed transition probabilities from homogeneous spontaneous activity, indicated as '0', to structured spontaneous activity, indicated as '1', and *vice versa* and we used them to determine the region of the parameter space where one can observe a dynamics close to that observe experimentally. We observe that, as J_E increases, the probability of state-emergence increases and the probability of state-decay decreases (Fig. 4.7 (A), the top row reports results from the noisy synaptic transmission model, the bottom row from the noisy feedforward drive model). Multiplying these probabilities yields an indicator that helps to identify a region where cortical states spontaneously emergence and decay (Fig. 4.7 (B)). Such region comprises values of J_E located around the boundary between linear and marginal phase (marked by the vertical black line), with the spread around the boundary increasing with the level of variability. Note that, interestingly, when the noise is in the synaptic process (top row), increasing the variability also influences probabilities.

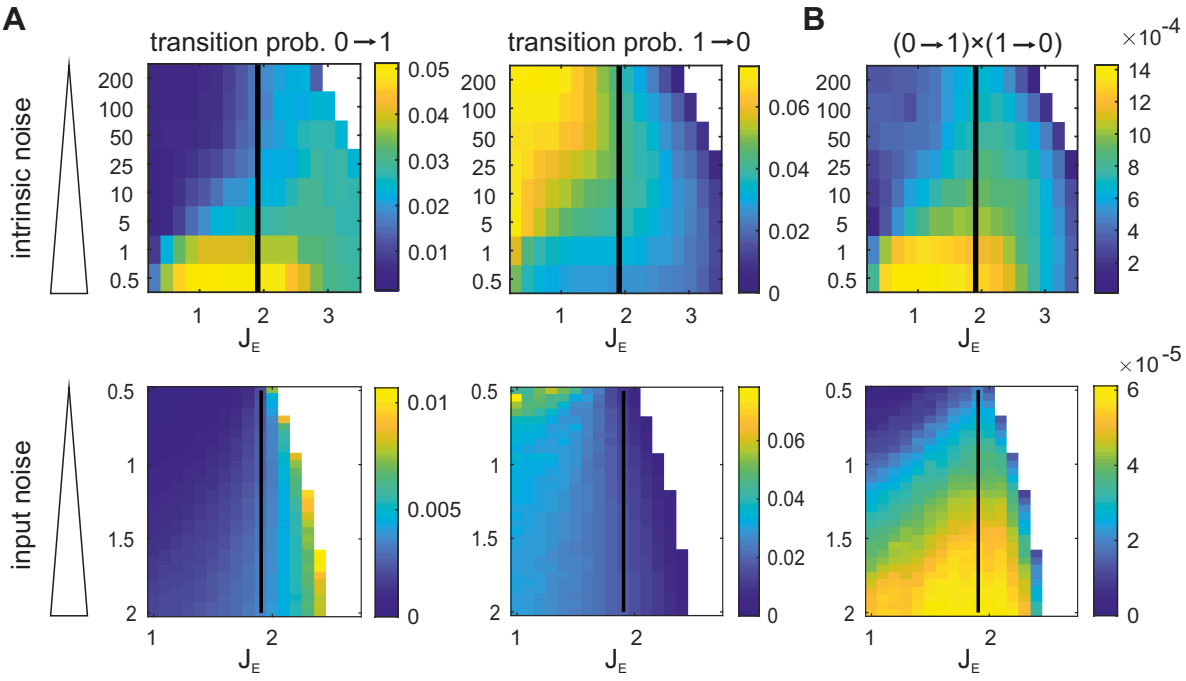


Fig. 4.7: Emergence and decay of spontaneous states. (**A**) Transition probabilities from random (0) to structured (1) spontaneous activity and from structured to random. (**B**) Indicator to detect emergence and decays of spontaneous states. Top and bottom rows show, respectively, simulations of Eq. (4.16) and (4.17), where we set $N_H = 20$, $N_P = 100$, $\sigma_I = 0.35$, $\sigma_E = 0.15$, $J_I = 2.5$, $\Lambda = 0.25$.

To further characterize the spontaneous states produced by our model, we also quantified the average duration and the average lateral spread of a state. Fig. 4.8 shows that spontaneous states tend to persist for longer times as the excitatory coupling strength gets larger. Considering together the range of mean and of the standard deviation of the state-persistence, our model is able to produce states whose duration is comparable to what has been observed experimentally, around 200 ms (O'hashi et al., 2017). Experiments also report that when states occur, they either

involve a few iso-orientation domains (in awake conditions) or the whole imaged area (under anesthesia). The lateral spread of the simulated states (quantified as the average number of hypercolumns in a state) is shown in Fig. 4.9 as a function of J_E and of the noise level. For both models, as we increase the coupling strength, the spontaneous states transition from a realistic regime in which they span between 1 and 5 hypercolumns to a state where a activity pattern spreads across the whole cortex (note that, in Fig. 4.9, 20 is the total number of simulated hypercolumns). The dependency shown in Fig. 4.9 does not change when we vary the length scale of long-range connections (except for the limit case of $\Lambda \approx 1$).

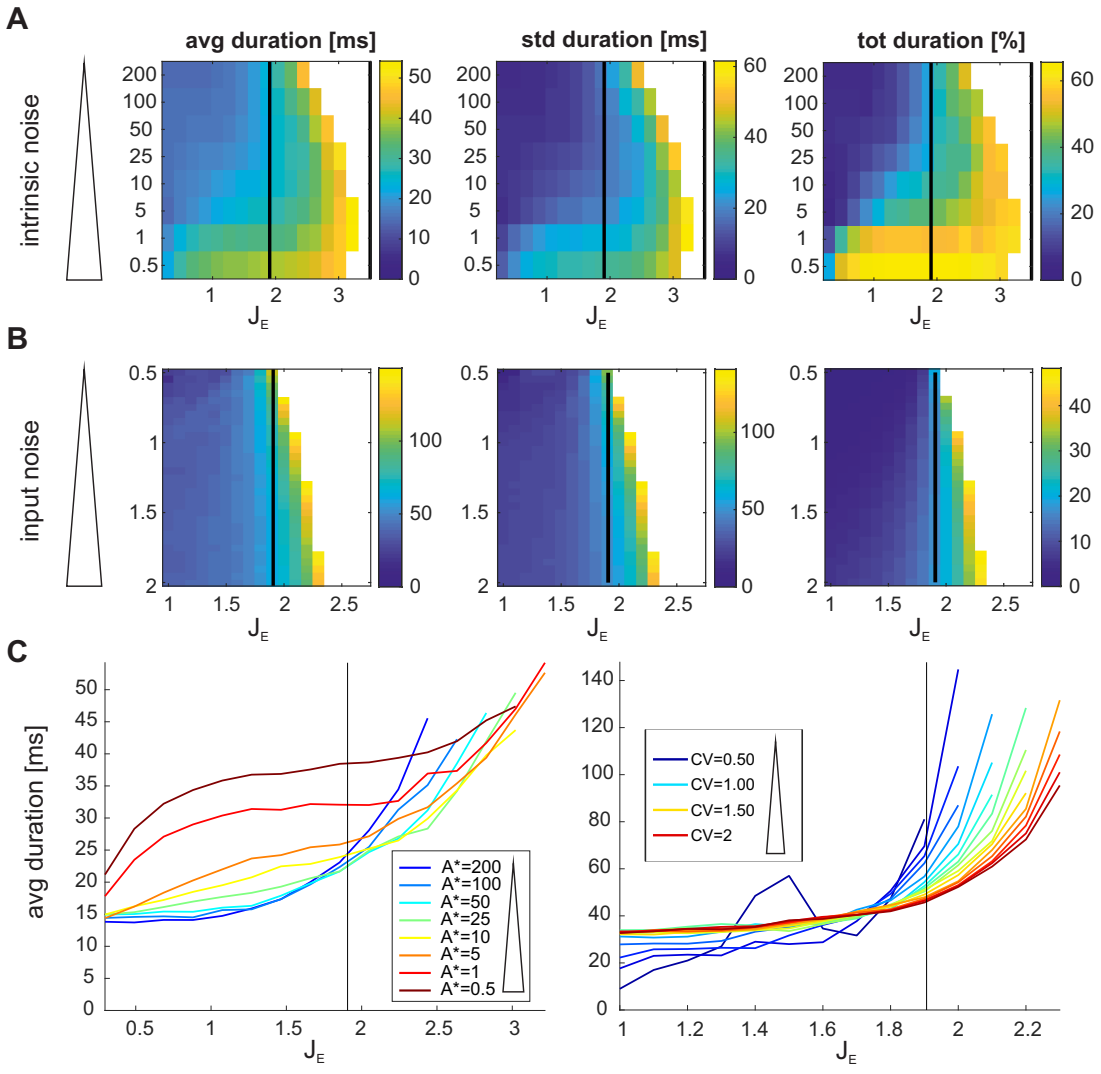


Fig. 4.8: Duration related statistics. Average duration of spontaneous tuned states, standard deviation of the duration and total time spent in a spontaneous states (computed as a percentage of the total simulated time) shown as a function of the noise level (y–axis, ordered from low (top) to high (bottom)) and the excitatory coupling (x–axis). Statistics are computed over approximately 100 spontaneous states, after stimulating Eq. (4.16) (**A**) and (4.17) (**B**), where we set $N_H = 20$, $N_P = 100$, $\sigma_I = 0.35$, $\sigma_E = 0.15$, $J_I = 2.5$, $\Lambda = 0.25$. (**C**) Average duration of spontaneous tuned drawn as curves (left contains simulations from Eq. (4.16) and right from Eq. (4.17)).

Another feature observed in experiments is the emergence of so-called mosaic states, that is patterns that are composed of states tuned to different orientations in different regions of the imaged area, instead of a unique coherent orientation. Our model also exhibits such mixed states (see Fig. 4.6). Quantifying their prevalence as a function of the spatial scale of the long-range interactions Λ (Fig. 4.10) reveals that the larger the extent of lateral coupling, the more coherent the states become – note that the actual values of the coherence measure depend also on the total simulated cortical area.

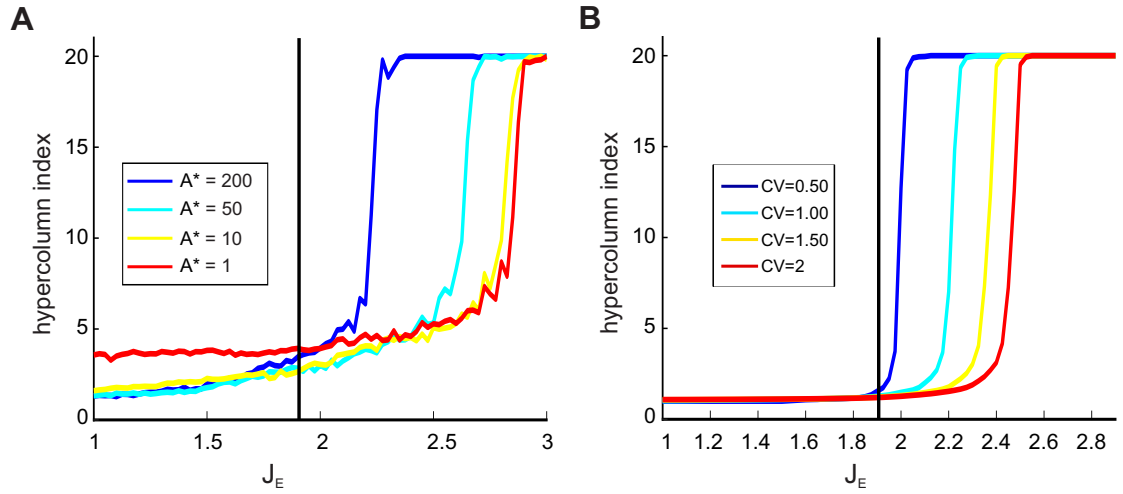


Fig. 4.9: Lateral spread of spontaneous states. Average number of hypercolumns that make up a spontaneous state computed as a function of excitatory coupling strength in the noisy synaptic transmission model (**A**), and in the noisy feedforward drive model (**B**) for different noise levels (black vertical line represents the boundary between linear and marginal phase). The remaining parameters were set as $N_H = 20$, $N_P = 100$, $\sigma_I = 0.35$, $\sigma_E = 0.15$, $J_I = 2.5$, $\Lambda = 0.25$.

Finally, to identify possible mechanisms underlying emergence and decay of states, we investigated the relative contribution of external and recurrent input to the populations within an hypercolumn. To do this, we averaged $I^{\text{ext}}(r,t)$ and $[W \star A](r,t)$ across the time-point when the activity profile starts to exhibit orientation tuning (see Methods) over approximately a thousand spontaneous states (Fig. 4.11). We found that states are triggered by the noisy input when it is, by chance, tuned. By looking at Fig. 4.11, we note that I^{ext} appears tuned to the emergent angle before the activity in the hypercolumn does. Moreover, this external tuned drive lasts, on average, for a period of 15 – 20 ms, consistent with the time constant of the system τ. Once a orientation-tuned state is initiated, it is actively maintained by the recurrent connectivity. A similar mechanism terminates the states, namely the arrival of external input to neural populations that represent a different orientation than the one currently active.

4.3 Discussion

4.3.1 Establishing the presence of spontaneous patterns

The quantitative analysis done to analyze the stochastic models relies on the procedure with which we detect spontaneous states. We opted for identifying localized activity-bumps from the circular variance, a standard measure of tuning when dealing with angular variables (Ringach

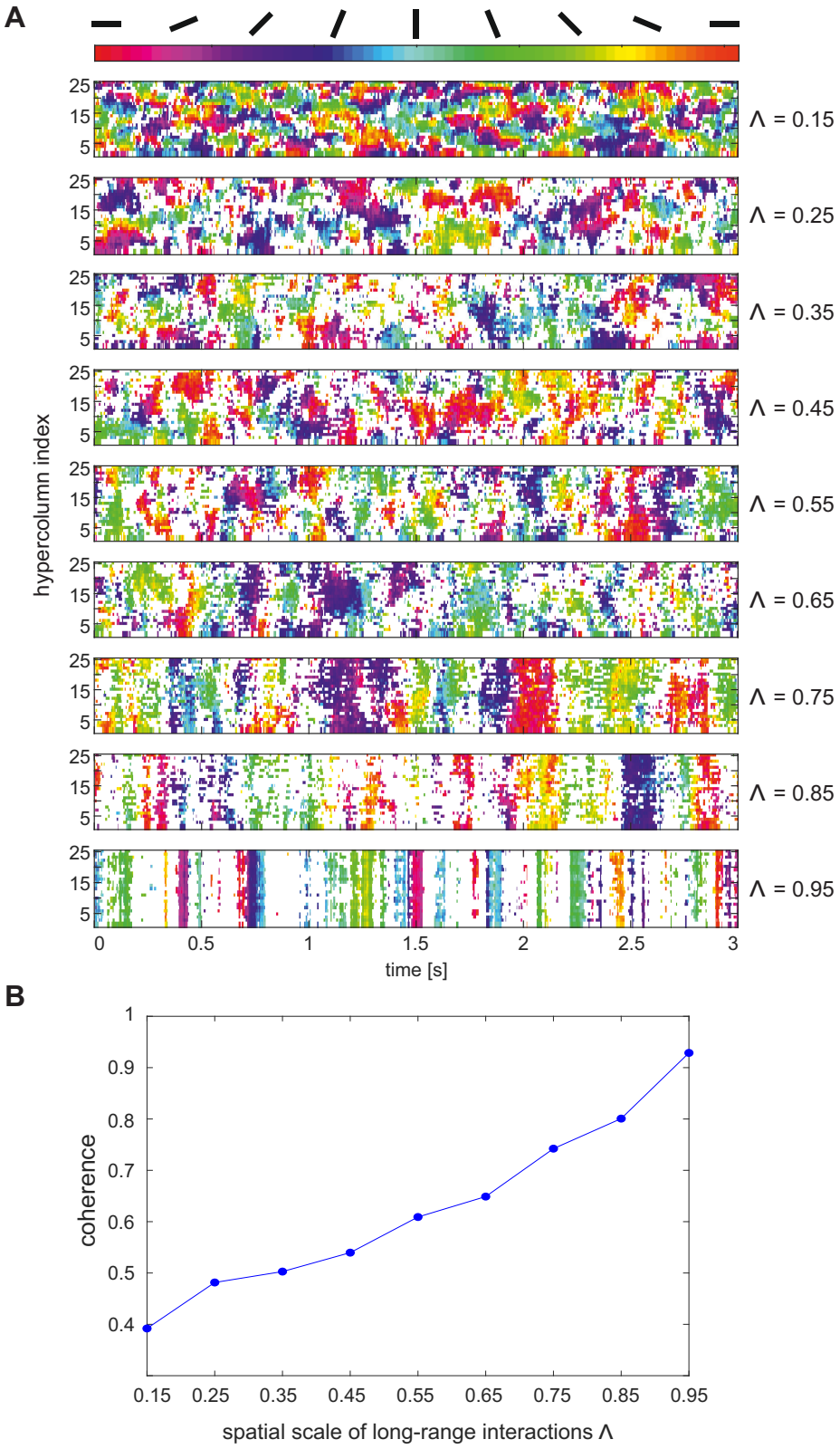


Fig. 4.10: Coherence. (**A**) Top: examples of emerging states for different values of Λ. As Λ increases, states become more coherent, i.e. we observe fewer mosaic states. (**B**) Bottom: Coherence of the states for different values of Λ. $N_H = 20$, $N_P = 100$, $\sigma_I = 0.35$, $\sigma_E = 0.15$, $J_I = 2.5$, $\Lambda = 0.25$

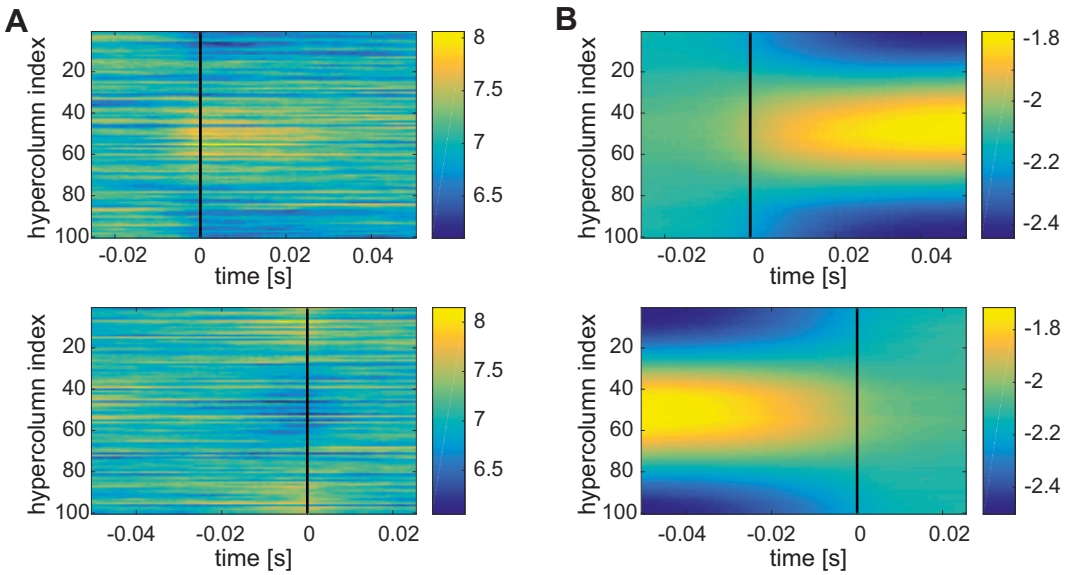


Fig. 4.11: (**A**) Average external and (**B**) recurrent input across the beginning (top row) and across the end (bottom row) of spontaneous tuned states. The average is computed over 370 states obtained stimulating the model described by Eq. (4.17) with the following parameters: $N_H = 20$, $N_P = 100$, $\sigma_I = 0.35$, $\sigma_E = 0.15$, $J_I = 2.5$, $\Lambda = 0.25$.

et al., 2002) which is, in addition, almost equivalent to the method used in (O'hashi et al., 2017), where they computed correlations between each image-frames to the best fitting orientation tuning template. However, the distribution of circular variance varies, to some extent, with all the parameters of the model, making it difficult to uniquely determine an appropriate threshold value, especially in a model corrupted by noise.

4.3.2 Implementation of a stochastic model

We observed that the introduction of noise in the dynamics is necessary to produce the empirically observed transitions between states, that is emergence, decay and a combination of smooth transition and abrupt switching. However, the question remains about what is the correct implementation of noise. We proposed two realizations, one that assumes that the state-transitions are caused by synaptic reliability and one that assumes that states' dynamics is governed by a noisy feedforward drive bearing some temporal correlation. Perhaps it would be appropriate to use a combination of the two mechanisms, since they are both present in cortex. An even more realistic implementation of the noisy input that V1 receives both from the lateral geniculate nucleus and from extra-striate areas would be to model an input with spatial correlations.

Another model that aimed at explaining spontaneous state-transitions (O'hashi et al., 2017) employed time-varying noise: to simulate ongoing activity the network was driven with i.i.d. noise equally distributed in a restricted interval, where the upper limit of such interval was modulated in a low-frequency range, reminiscent of slow wave activity observed across cortex. This achieved a good agreement with experimental results, because as the the upper limit of the noise increased from low to moderate to high, the transitions progressively went from absent to abrupt, to smooth.

4.3.3 Possible mechanisms underlying emergence and decay of states

Our model provides a possible explanation of the mechanisms that make cortical states appear and vanish. We propose that spontaneous oriented states are triggered upon receiving tuned input, persist because of the lateral connectivity, when excitation is strong enough to trigger a positive feedback-loop and terminate when either unspecific external input or external input tuned to a different orientation breaks the concentration of the activity.

4.3.4 Matching model results to experiments

The model offers a good qualitative match to the spatial features that are observed in imaging experiments: When states emerge they are tuned to a specific orientation and span several hypercolumns, resembling typical single-orientation maps and stochastic transitions to proximal as well as distant orientations are observed. Moreover the model predicts that increasing the length scale of long-range connections (parameter Λ) translates into a more coherent activity. Regarding the temporal aspects, the model is able to produce spontaneous states that persist for a duration similar to that observed in cortex. However, little discrepancies already exist between the reported data. For example (O'hashi et al., 2017) reports average duration values shorter than 200 ms, while (Smith et al., 2018) of about 1 s. Those discrepancies are most likely due to the different procedures used to define spontaneous events, but makes it hard to have a good quantitative comparison. We note here that state duration is influenced the time constant τ with which the activity decays and, in the model with the Gaussian input drive, also by the time constant of the noise, τ_η. A more precise quantitative match could probably be achieved if (i) more detailed statistical analysis were available from experiments and (ii) we realized a more biophysically plausible neural dynamics – for example implementing a spiking mechanism, synaptic delays or distinct excitatory and inhibitory neural populations.

4.4 Methods

As already explained in Section 4.2.3, the cortical space $\mathscr{C}$ was modeled with periodic boundary conditions, by considering the interval $\mathscr{C} = (0, 2N_H\pi)$. Note that $\mathscr{C}$ is composed by N_H fundamental periods of length 2π, each corresponding to an hypercolumn. This space was discretized with spatial resolution $\Delta x = \frac{2\pi}{N_P}$, which allows to denote the activity of a population at position $x_n = n\Delta x$, $n = 1, 2, \ldots, N_H \cdot N_P$ with A_n.

Numerical simulations of the differential equations that define the dynamics (Eqs. (4.2), (4.16) and (4.17)) were performed using an Euler integration scheme with time step $\Delta t = 0.1$ ms. The time constant τ was set to 20 ms while all the other parameters were varied as indicated in the Results.

To simulate Eq. (4.16), we approximated the instantaneous firing rate of population n with

$$\hat{r}_n(t) \approx A_n(t)\Delta t.$$

At each time step, we generated spikes $\rho_n(t)$ from a poisson distribution with rate $\hat{r}_n(t)$, which were realized as δ-functions.

To stimulate the stochastic drive in Eq. (4.17), we used the Euler-Maruyama scheme, that is, we uses the following discretization

$$I^{\text{ext}}(t + \Delta t) = I^{\text{ext}}(t) + \frac{\Delta t}{\tau_\eta}(\,\mu - I^{\text{ext}}(t)\,) + \sqrt{\frac{\Delta t}{\tau_\eta}}\sigma\xi(t), \quad \xi(t) \sim \mathcal{N}(0, 1).$$

4.4.1 Detecting a spontaneous orientation-tuned state

Using the discretization just described, at every point in time t the activity $A_n(t)$ is defined on a circular variable x_n. Its complex mean can thus be computed either averaging over the whole cortex

$$\zeta = \frac{\sum_{n=1}^{N_H \cdot N_P} A_n(t) e^{i x_n}}{\sum_{n=1}^{N_H \cdot N_P} A_n(t)} \in \mathbb{C} \tag{4.18}$$

or averaging over single hypercolumns

$$\zeta_j = \frac{\sum_{n=(j-1)N_P+1}^{jN_P} A_n(t) e^{i x_n}}{\sum_{n=(j-1)N_P+1}^{jN_P} A_n(t)} \in \mathbb{C} \tag{4.19}$$

for $j = 1, \ldots, N_H$. The argument of ζ and its absolute value are measures of location and spread of the activity. Specifically, the quantities

$$\theta(t) = \frac{\arg \zeta(t)}{2} \tag{4.20}$$

and

$$\theta_j(t) = \frac{\arg \zeta_j(t)}{2} \tag{4.21}$$

represent the orientations that spontaneously emerge either globally (Eq. (4.20)) or locally (Eq. (4.21)), while the quantities

$$z(t) = |\zeta(t)| \tag{4.22}$$

and

$$z_j(t) = |\zeta_j(t)| \tag{4.23}$$

represent how sharply tuned the states is, with values close to 0 indicating a flat profile in orientation space and values close to 1 indicating well concentrated data. We identified the presence of a spontaneous state in hypercolumn j at time t if $z_j(t)$ is above a certain threshold $z_{\text{thr}} = 0.1$. The value z_{thr} is chosen as the 99^{th} percentile of the distribution of $z_{\text{shuffled}}(t)$ computed after shuffling the activity over the spatial dimension.

This thresholding operation yields, for every simulated hypercolumn, a binary time series that can be used to compute further statistics. In particular, we defined the lateral spread of a spontaneous state as the number of adjacent hypercolumns that are above threshold at a particular time t and the duration of a state[1] as the average amount of time during which z_j is continuously above threshold.

To characterize the presence of mosaic states (i.e. pattern of activity containing states tuned to different orientations), we considered how similar the global tuning and the mean local tuning were, by computing the ratio

$$\frac{z(t)}{\sum_j z_j(t)/N_H}. \tag{4.24}$$

[1] To study how the average duration of a state depends on parameters, we excluded from the analysis all the states that were shorter than 5 ms, this duration being the sampling period typically used in experiments with VSDI.

This measure quantifies the level of 'coherence' of a spontaneous cortical state: When many hypercolumns are tuned to the same orientation, then a global orientation pattern emerges and the ratio in Eq. (4.24) is close to 1, since the numerator and the denominator have similar values. When many hypercolumns exhibit well-tuned states, but to different orientations (i.e. the denominator is close to one but the numerator is close to zero), then the ratio in Eq. (4.24) is low.

4.4.2 Simulations of the stochastic models

To identify the region of emergence and decay of spontaneous states, we integrated Eq. (4.16) and (4.17) varying the excitatory coupling strength and the noise, while holding the inhibitory coupling fixed ($J_I = 2.5$).

In particular, in the model where the variability is generated by unreliable synapses (Eq. 4.16), we chose the following constant input

$$I^{\text{ext}} = A^\star \cdot [1/\gamma + J_I - J_E] + b,$$

so that the dynamics would have $A^\star$ as fixed point, and we varied $A^\star$. The parameter $A^\star$ took values from the following list $[200, 100, 50, 25, 5, 1, 0.5]$ and J_E assumed 20 equidistant values in the range $[0.3, 3.4]$.

In the model where the variability is generated by a stochastic drive, we simulated Eq. (4.17) using as drift coefficient

$$\mu = A^\star \cdot [1/\gamma + J_I - J_E] + b,$$

so that the dynamics would have $A^\star$ as fixed point, and we varied σ. We fixed $A^\star = 5$ and we varied J_E between 1 and 2.7. We chose σ such that the coefficient of variation $\frac{\sigma}{\mu}$ would assume values between 0.5 and 2.

4.4.3 Transition probabilities

Transitions from a flat activity profile (state 0) to a sharply tuned activity profile (state 1) were computed by counting the number of times the following indicator function

$$F_j(t) := z_j(t) > z_{\text{thr}}$$

changed from x at time t to y at time $t + 1$, with $x, y \in \{0, 1\}$. To get a global measure, we averaged over the dimension j.

5 Evoking oriented percepts

5.1 Introduction

Many of the diseases leading to blindness often damage the retina or other ocular structures, leaving the remainder of the visual pathway still functional. A viable option to partially restore the sense of vision in patients with this type of acquired blindness is to insert artificial-visual signals directly into the brain. Recent advances in technology, in fact, make possible to create prosthetic devices capable of delivering intra-cortical stimulation to neural circuits with high spatial resolution and small current magnitude (Schmidt et al., 1996; Roelfsema et al., 2018). This goal of building a visual prosthesis dates back to the work of Brindley and Lewin (Brindley and Lewin, 1968) and Dobelle (Dobelle and Mladejovsky, 1974) who studied *phosphenes*, the perceptual sensations evoked by electrical stimulation of the occipital cortex, and is now pursued by many research groups (Lowery, 2013; Troyk, 2017).

Conventional electrical stimulation paradigms rely on the well-known retinotopic organization of primary visual cortex: By implanting an array of electrodes at different cortical locations one can ideally induce phosphenes across the whole visual field, with each electrode contributing one phosphene. However, concurrent stimulation of multiple cortical sites has resulted, so far, only in percepts of isolated phosphenes, that did not combine to form a coherent shape. One reason for this 'unnatural' perception, besides the distance between phosphenes, could be their round appearance, which most likely results from the broad activation of large population of neurons: given the fixed geometry of electrode array, microstimulation targets cells with all possible orientation preferences, resulting in an unspecific percept.

Moreover, the application of a similar paradigm does not take into account the current dynamical state of the brain, overriding - and possibly conflicting with - any already ongoing process. Both anatomically and functionally, the primary visual cortex is a highly structured network with abundant recurrent interactions that, even in the absence of a specific visual input, it is known to show spontaneous activation patterns that reflect functional connectivity, where iso-orientation domains are active either simultaneously or in close temporal proximity (Kenet et al., 2003; O'hashi et al., 2017; Omer et al., 2018; Smith et al., 2018).

In this chapter, we present a new stimulation-paradigm promising to achieve oriented percepts. This paradigm consists in monitoring the emergence, decay and nature of spontaneous orientation-tuned states that visual cortex encompasses and delivering an electric pulse when the cortex is in a desired state. In this way, we could use a weak modulatory current to induce spikes in neurons that are currently close to their firing threshold and have a preferred orientation similar to the orientation of the current visual image at the particular RF location. This would result in the percept of an elongated oriented feature, perhaps easier to combine with other oriented features into complex shapes. In particular, we use a structurally simple model of V1 (inspired from the one presented in Chapter 4) as a test bed to evaluate the proposed framework. We show that the activity of the model resembles spontaneous activity in V1,

i.e. the network randomly switches between different states tuned to particular orientations, separated by periods where activation is more stochastic and less coordinated. We calibrate the model to a physiologically realistic operating point and, in this idealized setting, we conduct a feasibility study, investigating in particular the relations between stimulation amplitude, temporal resolution and specificity of the percept. Last, we delineate a strategy to evoke complex percepts.

5.2 Results

5.2.1 Model description

To model ongoing activity in primary visual cortex under the effects of intracortical stimulation, we consider a network of integrate and fire neurons. In particular, we simulate the activity of N_H hypercolumns arranged in a one-dimensional space (where we impose periodic boundary conditions), each composed by N neurons characterized by an orientation preference

$$\theta_n^{\text{pref}} = \frac{\pi n}{N} \pmod{\pi}, \text{ with } n = 1, \dots, N \cdot N_H. \tag{5.1}$$

The sub-threshold dynamics of neuron n is described by its membrane potential

$$\tau \frac{dV_n}{dt} = -V_n(t) + V_{\text{rest}} + x_n(t) + \eta_n(t) + I^{\text{pulse}}(t). \tag{5.2}$$

Whenever the membrane potential reaches a threshold V_{thr} a spike is emitted and V_n is reset to the resting potential V_{rest}.

The term $x_n(t) = x_n^E(t) - x_n^I(t)$ represents the synaptic currents, whose temporal evolutions read

$$\tau_E \frac{dx_n^E}{dt} = -x_n^E + \tau_E \sum_{k=1}^{N} W_{nk}^{\text{long}} \rho_k(t) \tag{5.3}$$

$$\tau_I \frac{dx_n^I}{dt} = -x_n^I + \tau_I \sum_{k=1}^{N} W_{nk}^{\text{local}} \rho_k(t). \tag{5.4}$$

Here, ρ_k indicates the spikes from presynaptic neurons k, W_{nk}^{long} and W_{nk}^{loc} the excitatory and inhibitory synaptic weights, and τ_E, τ_I are two time constants. The synaptic connections are defined as in Chapter 4 (Eqs. 4.6 and 4.5), namely

$$W_{nk}^{\text{local}} = -J_I \frac{1}{\sqrt{2\pi}\sigma_I} \exp\Big(- \frac{|r_n - r_k|^2}{2\sigma_I^2} \Big), \tag{5.5}$$

$$W_{nk}^{\text{long}} = J_E \frac{1-\Lambda}{1+\Lambda} \frac{1}{\sqrt{2\pi}\sigma_E} \sum_{j=-\infty}^{\infty} \exp\Big(- \frac{|r_n - r_k - 2j\pi|^2}{2\sigma_E^2} \Big) \Lambda^{|j|}, \tag{5.6}$$

where the variable r_n denotes the position of neuron n in cortex. Since we assume a cortical space with periodic boundary conditions, discretized with a spatial resolution of $\Delta r = 2\pi/N$, we have $r_n = n\Delta r$. The term η_n is a temporally correlated Gaussian noise with mean μ, variance σ^2 and time constant τ_η (see Section 5.4.1 for details on the implementation) and it models thalamic input as well as feedback from other areas. The parameters μ and σ are chosen such that the noise level is just big enough to bring neurons to fire with low rates.

The term I^{pulse} represents an artificial electric input, later referred to as a 'pulse'. Here we model an electric pulse as an external step current with amplitude Δa and a fixed 1 ms duration.

5.2.2 Combining ongoing dynamics with electrical pulses

The activity produced by the model resembles ongoing activity in primary visual cortex: the network exhibits low firing rates and a high degree of variability (Ringach et al., 2002); moreover, bursts of spikes localized in space, spanning a few hypercolumns tend to emerge spontaneously (Fig. 5.1 (A)). Analyzing the orientation preference of the neurons participating to the spontaneous state (Section 5.4.2) reveals a spatio-temporal pattern (Fig. 5.1 (B)) similar to what has been observed experimentally in cat and monkey ongoing activity (Kenet et al., 2003; O'hashi et al., 2017).

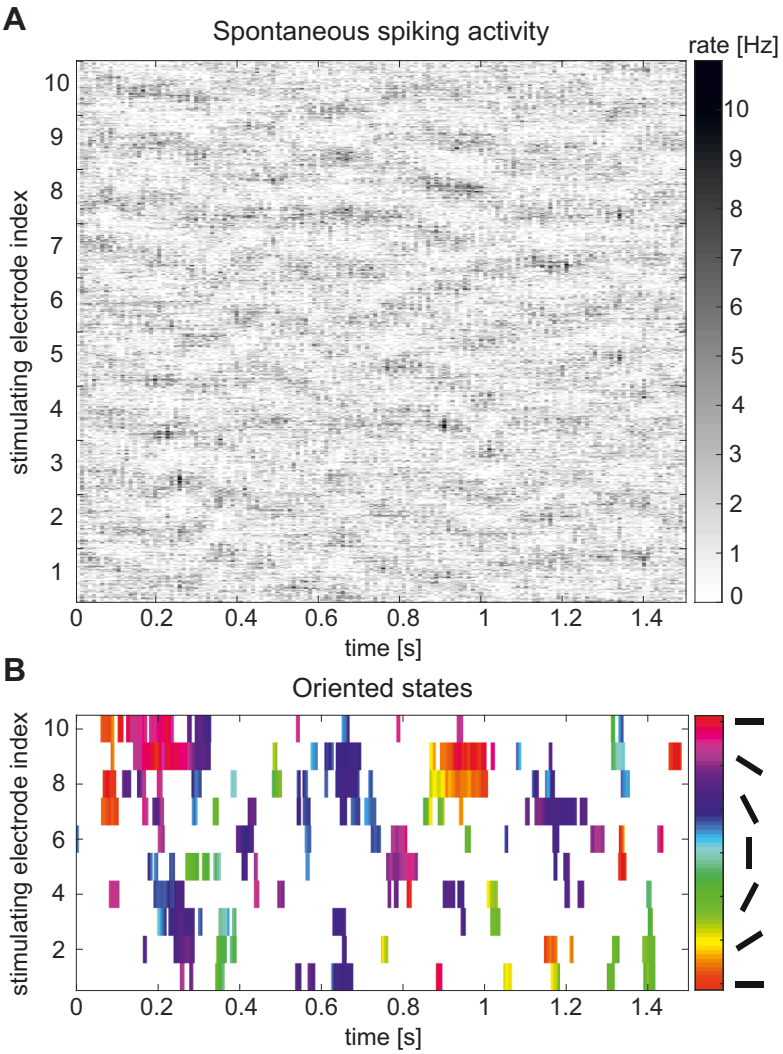


Fig. 5.1: Spontaneous states. (**A**) Spontaneous spiking activity resulting from a simulation with no electrical stimuli. To improve visibility, spikes are cumulated over groups of 20 neurons and over 10 ms time-intervals. The average firing rate is approximately 7 Hz. (**B**) Spontaneous oriented states corresponding to the activity shown in (**A**). Orientations are colored as indicated in the colorbar.

Our idea to implement a stimulation protocol to evoke oriented percept is the following:

- We assume we can observe the spontaneous activity in a patch of primary visual cortex. In this thesis, for the sake of simplicity, we imagine that we can stimulate from

N_H electrodes arranged in a grid and placed at a distance d_E from each other which corresponds to the average distance between orientation hypercolumns' centers. Each stimulation-electrode is surrounded by a neighborhood of N recording electrodes, each sampling a signal from a neural population with orientation preference ϕ_n^{pref}, with $n = 1, 2, \ldots, N$.

- For each electrode k, we specify a target orientation ϕ_k^{target} that we want to elicit.
- In an online fashion, we monitor the activity at each electrode site. When the activity at one position k is sufficiently tuned, i.e. above a certain 'specificity threshold' z_{thr}, we interpret it as a 'proper' network state and we determine its orientation ϕ_k^{spont} (see Section 5.4.2 for details).
- When, at one particular position, we detect a state whose orientation is sufficiently close to the target orientation ($|\phi_k^{\text{spont}} - \phi_k^{\text{target}}| < \Delta\phi$), we deliver a local electrical pulse through electrode k.
- After giving an electrical stimulus, we wait for τ_p seconds before allowing to deliver a second pulse to the same location.

Before proceeding with artificial stimulation of the network, we want to explain the reasons why we expect our protocol to work and which disadvantages it might have. The basic idea is quite straightforward: in a tuned state, some neurons fire with higher rates and thus their membrane potentials are on average closer to the firing threshold (Fig. 5.2). Giving an electric pulse of the right magnitude makes neurons close to the threshold fire action potentials, while neurons further away from threshold stay silent. The activated neurons signal the presence of an edge with orientation close to the desired orientation to downstream areas, thus inducing the intended percept. In this view, the amplitude of the pulse directly controls the probability of detecting the evoked percept. At the same time, however, we expect that if the amplitude of the pulse gets too large, it will bring a greater number of neurons closer to their firing threshold, rendering the orientation of the percept less discernible. We thus expect a trade-off between strength of the elicited percept and its specificity.

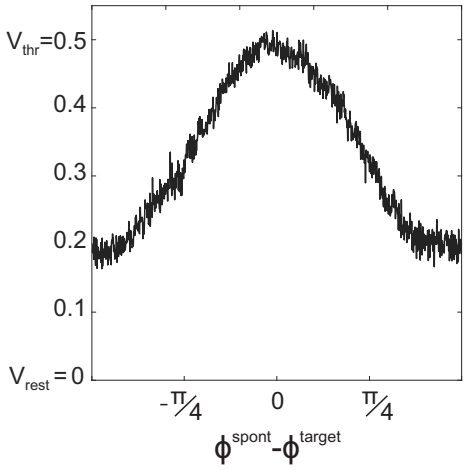


Fig. 5.2: Average values of membrane potential of the neurons in an hypercolumn during a spontaneous state. This distribution is obtained by averaging over ≈ 100 states.

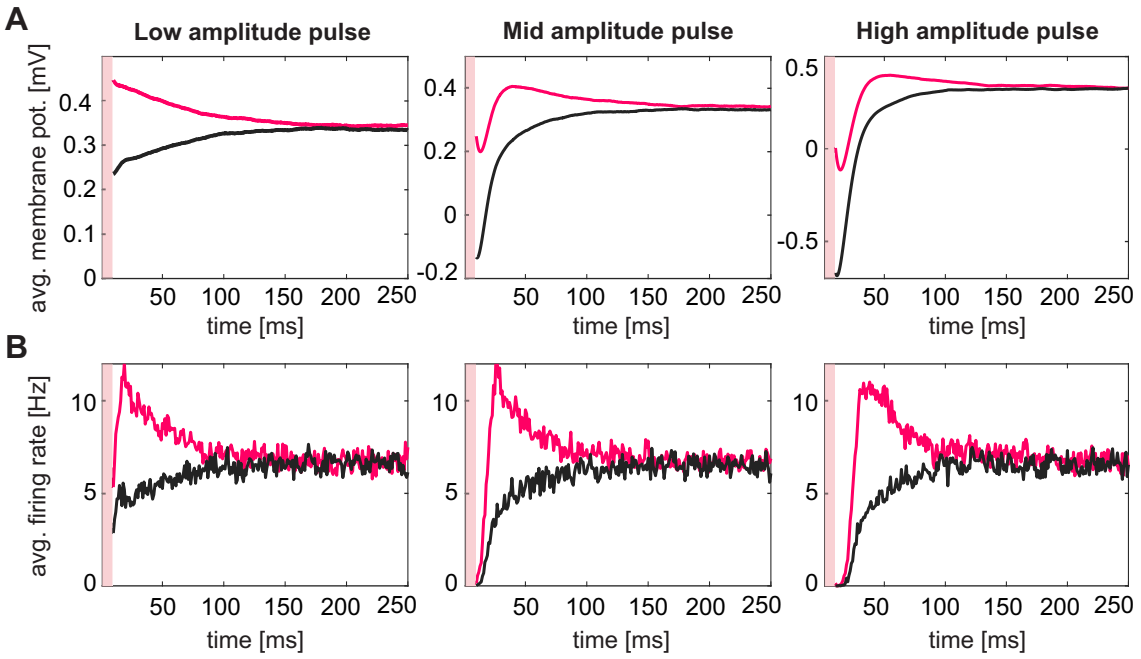


Fig. 5.3: Average time course of the membrane potential (**A**) and of the firing rate (**B**) after a pulse is delivered, presented separately for the neurons expressing the target orientation (red curves) and for the neurons expressing the orthogonal orientation (black curves), for three different pulse-amplitude values. Note that the initial 1 ms period when the electrical stimulation is delivered is blanked out.

5.2.3 Feasibility study

Detecting how long the effect of a single pulse lasts is important to establish what is the minimal time interval for the system to recover. In Fig. 5.3 we report the average time course of the membrane potential (panel (A)) and the average firing rate (panel (B)) following a pulse. The pink and gray curves are computed from neurons representing, respectively, the target orientation and the orthogonal orientation for a low, medium and high pulse amplitude. We see that between 100 – 200 ms after a pulse, pink and black curves collapse onto each other, demonstrating that the effect of electrical stimulation has disappeared (notice, in particular, that the firing rate settles back around 7 Hz, the spontaneous rate).

The average effect on the activity after receiving electrical stimulation is shown in Fig. 5.4 (A) for various values of Δa. Here, we included the activity following a zero-amplitude pulse (i.e. pure spontaneous activity), as a comparison. The distribution of firing rate indicates that the electrical stimulation enhances the tuning of the network state prior to the pulse without inducing an orientation shift. The rates reported in Fig. 5.4 (A) are calculated over a brief period of 10 ms right after an electrical pulse has been delivered: the high values on the vertical axis corresponds to a few spikes occurring almost synchronously. This suggests that the activity caused in V1 by the stimulation would have a significant impact on subsequent processing areas and, ultimately, on perception. Moreover we notice that, as the pulse amplitude increases, the activity has a greater offset, indicating a lower degree of specificity. Both of these observations are quantified in Fig. 5.4 (B), where we reported the average rate and the average specificity as a function of the pulse amplitude.

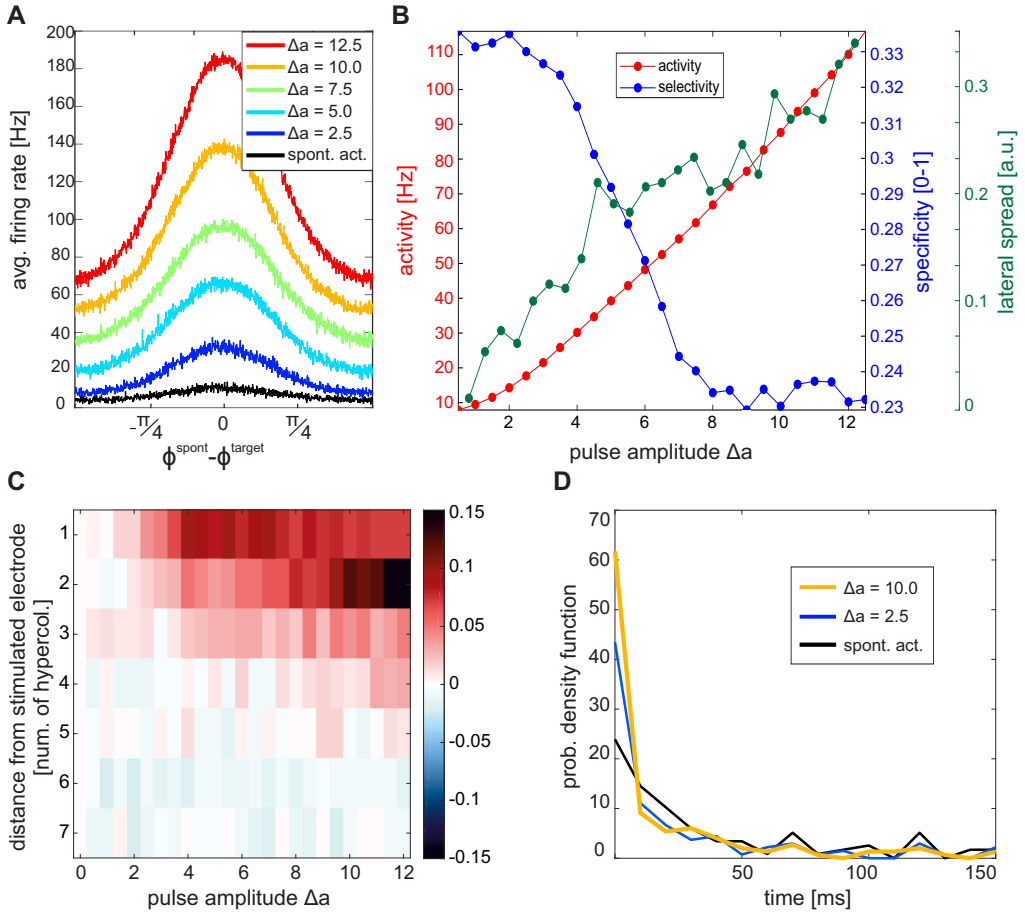


Fig. 5.4: Effects of varying pulse strength. (**A**) Average firing rate of the neurons in an hypercolumn following electric pulses of various amplitude. (**B**) Trade-off curve between the strength of evoked percept and its specificity. (**C**) Change in average number of oriented states observed up to 250 ms after delivering a localized pulse as a function of pulse strength and distance from the delivery site with respect to the number of states that would be observed at that particular distance during spontaneous activity. (**D**) Distribution of inter-state time intervals for spontaneous activity (black curve), low (blue curve) and high (yellow curve) pulse strengths.

An important aspect to consider, if we want to determine the optimal pulse strength to produce a localized oriented percept, is that electrical stimulation might not remain confined to the region surrounding the electrode tip. To investigate how the effect of a pulse delivered at one particular position spreads to neighboring locations, we monitored all electrodes k, selecting the same target orientation $\phi_k^{\text{target}} = \phi^{\text{target}}$, but we stimulated only one of them, say k_0. We then quantified how a pulse delivered at position k_0 increased the probability of observing a spontaneous state with orientation $\phi_k^{\text{spont}} = \phi^{\text{target}}$ at all possible positions k. For each position k and for each amplitude level Δa, we computed the average number of oriented states observed after a localized pulse delivered at k_0 and subtracted from it a baseline level given by the number of states that would be observed at that particular distance even in the absence of electrical stimulation (remember that Eq. (5.2) produces activity pattern that have pronounced spatio-temporal correlations). The results is shown in Fig. 5.4 (C) as a function of the amplitude Δa and the distance to k_0. We observe an increase in probability of finding a spontaneous

tuned state with respect to the baseline (red shades) for locations close to k_0 by less than 3 electrode-positions, strongly modulated by the pulse amplitude. For distances greater than 4 electrode-positions we observe a mild decrease in the probability of finding a spontaneous tuned state with respect to the baseline (blue shades)[1]. As the pulse amplitude increases, not only the probability of a pulse to induce a spurious state increases, but also the time window in which this happens gets shorter. In Fig. 5.4 (D) we show the distribution of the inter-state time intervals, pooling the data for short distances (i.e. $|k - k_0| = 1$). The probability of observing shorter time intervals is higher in the presence of stimulation with respect to the pure spontaneous activity, with appreciable differences between weak and strong stimulation. If, after a pulse, the induced orientation-tuned state spreads to other hypercolumns, then this also should be taken into account in the trade-off mentioned above. The total lateral 'spill out' of activity caused by a pulse of a given amplitude was quantified as the sum of absolute change in the number of spurious spontaneous states over all observed distances and it is shown in Fig. 5.4 (A) (green curve).
An example of how a simple oriented percept might look is depicted in Fig. 5.5.

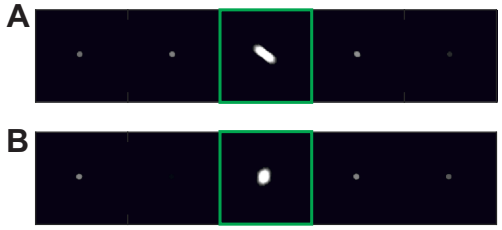


Fig. 5.5: Simple oriented percept. (**A**) Oriented artificial percept evoked by delivering a pulse during a sharply tuned state. (**B**) Round phosphene evoked by delivering a pulse during a period of unstructured spontaneous activity.

To get closer to the ambitious goal of vision restoration, besides producing oriented percepts, a stimulation protocol should at least be capable to evoke simple figures such as, for example, elementary geometrical shapes or letters. Since the mentioned figures are composed of combinations of oriented features, the electrical stimulation paradigm proposed should be able to produce them, provided that oriented states occur with a sufficiently high temporal resolution. We then investigated how often a particular target orientation occurs as a function of the specificity threshold z_{thr} and the precision required on the evoked percept $\Delta\phi$. We call the time to wait to obtain a certain orientation to occur spontaneously t_ϕ^{wait} and we consider $t^{\text{wait}} = \langle t_\phi^{\text{wait}} \rangle_\phi$, since the symmetry in our coupling matrices guarantees that every orientation occur spontaneously with the same probability. We find, as intuition would suggest, that t^{wait} increases with increasing precision and decreases by increasing the specificity threshold. In particular, we were able to recognize the following dependencies

$$t^{\text{wait}}(\Delta\phi) = \frac{A_1}{\Delta\phi} + A_0, \tag{5.7}$$

$$t^{\text{wait}}(z_{\text{thr}}) = B_0 \exp(B_1 z_{\text{thr}}), \tag{5.8}$$

where A_0 and A_1 depend on z_{thr}, while B_0 and B_1 on $\Delta\phi$.

5.2.4 Towards more complex percepts

As an application of the stimulation paradigm we described, we simulated how we could evoke a more complex image than a single oriented edge, such as the espresso cup shown in Fig. 5.6.

[1] We expect that an extensive parameter scan would reveal that these numbers are determined by Λ

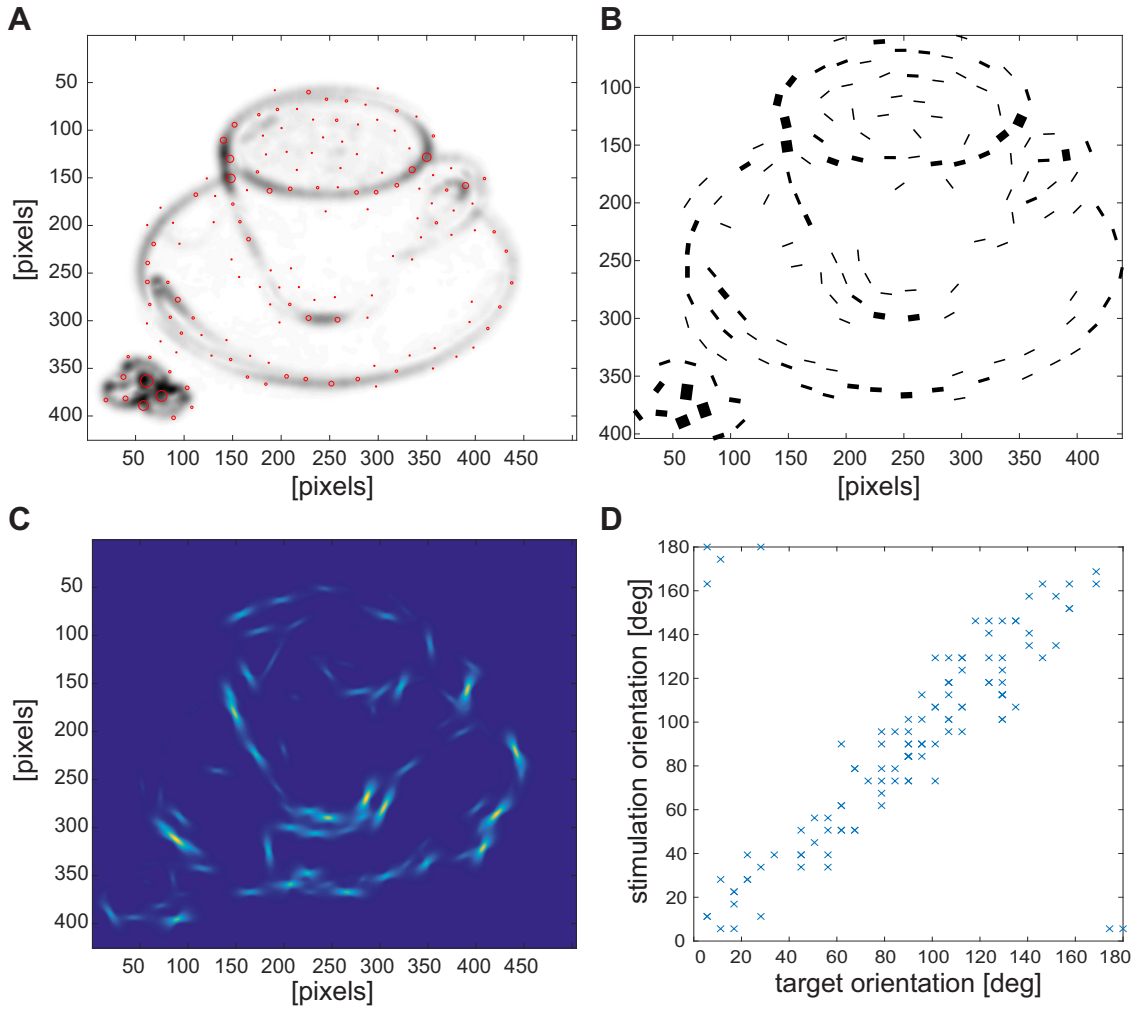


Fig. 5.6: (**A**) Filtered image where sampling points showing substantial tuning superimposed. (**B**) Local orientation content of the image. (**C**) Elicited percept. (**D**) Agreement between target and elicited orientation at a particular stimulation site.

To do this, we first extracted the orientation content at each location of the image, as shown in Fig. 5.6 (B) (Section 5.4.3). Assuming a specific arrangement of the electrodes and a retinotopic correspondence with regions of the image, we identified the stimulation parameters – mainly stimulation sites target orientations (Fig. 5.6 (A)). As explained in Section 5.2.2, we realized stimulation by integrating for a period of 1 second Eq. (5.2) and by delivering localized single pulses as soon as a spontaneous state tuned to the target orientation was detected. The artificially induced percept should ideally presented in a movie showing the how the effect of electrical pulses combined with the ongoing activity. In Fig. 5.6 (C), we present a single frame of such a movie. For assessing success of the stimulation protocol, we extracted activity from the monitored hypercolumns in a 5 ms time window from pulse onset. Orientation tuning of the activation profile corresponds well with the target orientations, as shown in the scatter plot Fig. 5.6 (D).

5.3 Discussion

The parameter regime in which the current model of ongoing activity is used – a regime of low firing rate whose activation looks sparse and irregular and occasionally transition stochastically in more organized patterns resembling local orientation maps – was tuned by hand, with knowledge gained from the linear stability analysis performed in Chapter 4. It remains for future work to investigate the models' robustness against change in parameters. The intuition inherited from previous models and several simulations that are not shown here, suggest that increasing the standard deviation σ of the noisy input or the excitatory coupling strength J_E results in an increase of the average specificity (or tuning) of the oriented states, but often ends up in an unrealistic behavior of runaway activity, where the network gets stuck in one pattern. However, we don't know precisely how close the network of integrate and fire neurons is to the rate-model analyzed in the previous chapter, since their non-linear mechanisms are different from each other. One important theoretical step would be to verify whether all the properties valid for the mean-field model (found with stability analysis and numerical investigation) still hold, especially the partition of the parameter space in linear, marginal and divergent phase and the relation between the noise and the probability of state detection.

For this to be a realistic testbed to study the effects of electrical stimulation in cortex, two improvements are necessary. First, we would need a more complex architecture of the network, to simulate a two-dimensional cortical sheet – but note that, despite being a simple conceptual change, such implementation potentially hides many numerical difficulties. Second, we would need to introduce a term that models the spatial spread of current that even local electrical pulses induce.

5.4 Methods

5.4.1 Numerical simulations and parameters

Numerical simulations of Eq. (5.2) were performed using an Euler integration scheme with time step $\Delta t = 0.1$ ms.

The feedforward inputs η_n in Eq. (5.2) are defined as an Ornstein-Uhlenbeck processes with drift and diffusion coefficients given by μ/τ_η and $\sigma/\sqrt{\tau_\eta}$ and are independent for each neuron. Omitting the index n for simplicity, we have

$$d\eta_t = \frac{1}{\tau_\eta}(\,\mu - \eta_t\,)dt + \frac{\sigma}{\sqrt{\tau_\eta}}\,d\xi_t, \tag{5.9}$$

where ξ_t is a stochastic process whose increments are statistically independent and follow a Gaussian distribution. Equation (5.9) was integrated using the Euler-Maruyama method, using the following discretization

$$\eta(t+\Delta t) = -\eta(t) + \frac{\Delta t}{\tau_\eta}(\,\mu - \eta(t)\,) + \sqrt{\frac{\Delta t}{\tau_\eta}}\sigma\xi(t), \quad \xi(t) \sim \mathcal{N}(0,1). \tag{5.10}$$

The parameters that define the coupling matrices and the dynamics were held fixed, while the parameters related to electrical stimulation were varied (values are shown in Table 5.1). The other parameters were varied as indicated in the Results section.

Table 5.1: Parameters used in simulations

Model's parameters

τ	τ_E	τ_I	Λ	J_E	J_I
15ms	2ms	4ms	0.1	4.875	6.250
σ_E	σ_I	μ	σ	τ_η	τ_{conv}
0.15	0.35	0.6mV	0.8mV	50ms	5ms

Fig. 5.1 and 5.2

N_H	N_E	N	N_S	Δa
10	10	1000	1000	0

Fig. 5.3

N_H	N_E	N	N_S	$\Delta a_{\text{low,mid,high}}$	$\Delta\phi$
20	10	1000	1000	$0, 55, 77$	30°

Fig. 5.4

N_H	N_E	N	N_S	Δa	$\Delta\phi$
15	10	1000	1000	24 values from 0 to 12.5	30°

5.4.2 Detecting an orientation-tuned state

To detect the emergence and decay of spontaneous, orientation-tuned states, the same procedure as described in Chapter 4 was used. The spiking activity ρ_n is converted to a continuous firing rate $r(t)$ by means of a convolution with an exponential kernel

$$r_n(t) = \int_0^{-\infty} \rho_n(t-y)\exp(-y/\tau_{\text{conv}})dy,$$

with $\tau_{\text{conv}} = 15$ ms. Then the local complex average for every hypercolumn/electrode j is computed as in Eq. (4.19), that is

$$\zeta_j(t) = \frac{\sum_{n=(j-1)N_P+1}^{jN_P} r_n(t)e^{2i\theta_n}}{\sum_{n=(j-1)N_P+1}^{jN_P} r_n(t)} \in \mathbb{C}$$

for $j = 1, \ldots, N_H$. The hypercolumn j is said to be in a spontaneous oriented state if

$$z_j(t) = |\zeta_j(t)| > z_{\text{thr}},$$

with the threshold value set to $z_{\text{thr}} = 0.3$, and its orientation is

$$\phi_j^{\text{spont}}(t) = \frac{\arg\{\zeta_j(t)\}}{2}.$$

5.4.3 Reconstructing the evoked percepts

For encoding a given image $\boldsymbol{s}$, a region of interest $[x_{\text{frame}}, y_{\text{frame}}]$ was first extracted from the full picture. The image was then converted to gray scale and filtered using a filter bank of $N_{\text{ori}} = 8$ oriented, complex Gabor patches representing angles between 0 and π, given by

$$g(x, y, \theta) = \exp\left(-\frac{x^2+y^2}{2\sigma_g^2}\right)\exp\left(2\pi i\frac{x'}{\lambda_g} + i\psi\right)$$

where

$$x' = x\cos(\theta) + y\sin(\theta)$$

with size $\sigma_g = 5$ pixels and spatial period $\lambda_g = 8$ pixels. The complex Gabors served to provide quadrature filter pairs with phases shifted by $\pi/2$, $g_{\sin} = \Im\{g\}$ and $g_{\cos} = \Re\{g\}$, for obtaining a phase-independent estimate of local orientation for each image pixel

$$E(x, y, \theta) = \sqrt{(g_{\sin} \star \boldsymbol{s})^2 + (g_{\cos} \star \boldsymbol{s})^2}.$$

To avoid boundary effects, a border of width $2\sigma_g$ was removed from the filtered image before further processing.

In a next step, locations for electrical stimulation were determined. We assumed that the stimulation array would have a hexagonal arrangement of electrodes with jitter introduced by local distortions in the retinotopic mapping in cortical space. The locations were generated by Matlab code previously used in the context of psychophysical studies for distributing localized stimuli homogeneously over visual space (Ernst et al., 2012).

At each location, we assessed orientation selectivity and total filter activation relative to the maximum filter response to determine sites near salient, oriented, local image features. Only sites with absolute orientation selectivity above $z_{\text{thr}} = 0.2$ and relative activation above $a_{\text{thr}} = 0.02$ were scheduled for stimulation with a desired target orientation given by the argument of the (complex) selectivity value.

For probing stimulation, we simulated a one-dimensional network of $N_H = 1024$ hypercolumns each consisting of $N = 1024$ integrate-and-fire neurons (Eq. (5.2)). Electrode sites were mapped to neuron indices randomly according to the following procedure: the N_H hypercolumns were partitioned into $N_H/2$ chunks of two consecutive hypercolumns. From each chunk, we selected N consecutive neurons, starting from a randomly selected neuron index in the first half of the chunk. This selection provided us with a full hypercolumn for state monitoring and stimulation (N neurons), which, however, started with a random orientation preference. The aim of this procedure was to increase biophysical realism of the stimulation paradigm, by having the target orientation at different spatial locations (i.e., close to the border or close to the centre) within the ensemble of neurons affected by a stimulation pulse. At the same time, this setup also constrained the potential number of stimulation sites to $N_H/2$ for the procedure performing hexagonal placement. In the end, for the example reported in Fig. 5.6, we scheduled 150 sites for electric stimulation. We simulated a one-second time period of spontaneous activity with targeted electric stimulation. For facilitating data handling, we re-binned output spikes reducing orientation hypercolumns from 1024 to 32 columns, and time resolution to 1 ms.

For visualizing the (potentially) induced spatio-temporal percept, we did not directly use the activation profiles obtained from the simulation. Instead, we performed inference with the Zhu and Rozell (2013) equations on network activity, using a dictionary derived from the average activation profile obtained by superimposing the single activation profiles centered around their evoked orientation. In detail, we performed reconstruction using the differential equation[2]

$$\tau_h \dot{\boldsymbol{h}} = -\boldsymbol{h} + \Phi^\top \boldsymbol{s} - (\Phi^\top \Phi - \mathbb{I})\boldsymbol{a}$$
$$\boldsymbol{a} = [\boldsymbol{h} - \lambda_a]^+$$

[2]Compare to Eq. 3.15 with $C = 0$.

with a time discretization of $dt = 0.002$ ms and time constant $\tau = 0.020$ ms, using a sparseness constant $\lambda_a = 1$. Hence, for every activation profile of 1 ms duration, we performed 500 reconstruction steps.

For rendering the percept, reconstructed activity vectors over orientation space were positioned at the original sampling locations within the two-dimensional image boundaries, convolved with a filter bank of N_{ori} elongated Gaussian profiles (with long- and short-axis set resp. to 16 and 2 pixels), and finally averaged over orientation dimension. Temporal evolution of the percept over the 1 s simulation period could be visualized by rendering the reconstructed frames as a movie, hereby enhancing the representation for visual inspection through low-pass filtering with an exponential kernel with time constant $\tau_{\text{decay}} = 10$.

6 Conclusion

6.1 Summary

The broad purpose of this thesis was to build a theoretical setup to model the introduction of artificial percepts into visual cortex. To build such a model, we investigated two different aspects of visual information processing: contextual modulations and the spontaneous emergence of oriented cortical states.

In Chapter 3 we investigated how spatially extended natural scenes are encoded in the early visual cortex and in particular how the activity of individual neurons representing information from localized regions of the visual field is integrated to form coherent representations of stimuli. To do this, we choose to study non-linear modulations of neural responses observed when localized stimuli are placed in a broader spatial context known as 'non-classical receptive field phenomena'.

Our model belongs to a class of normative models based on the sparse-coding hypothesis, which states that each given stimulus should be efficiently encoded using only a small number of cells. It is not just an instance of sparse-coding, but it constitutes an extension of the classical framework. Indeed, the generative model of natural images we proposed allows to encode simultaneously multiple discrete regions that form a scene. The encoding is realized by explicitly introducing a set of variables with the purpose of capturing spatial dependencies among oriented features at distant locations.

Applying a computational coding principle (i.e. optimizing accuracy and sparseness of stimuli representation) in addition to an anatomical constraint (i.e. imposing that neurons have direct access only to small portions of the visual field) is sufficient to account for a number of heterogeneous features: (i) deriving a realistic neural dynamics realized by a network of populations, (ii) learning a set of connectivity structures conform to anatomical findings, (iii) replicating with a uniform explanation experimental finding about non-classical receptive field phenomena, some of which previously unexplained.

In the second part of this thesis we investigated which mechanisms regulate spontaneous activity in primary visual cortex (Chapter 4) and how spontaneous activity could be combined with a electrical stimulation to generate artificial percepts (Chapter 5).

We defined a structurally simple model and, performing linear stability analysis, we studied the patterns of activity that the model exhibited, identifying conditions and parameters under which biophysically realistic cortical states emerged. We then introduced noise in the system to allow spontaneous pattern formation and decay, and we quantified the stochastic dynamics of spontaneous states in terms of emergence and decay probabilities and average duration.

The connectivity of the network was picked to mimic what we obtained from the optimization process in the sparse coding model. Synaptic interactions were thus characterized by two terms: an inhibitory coupling term, acting locally and an excitatory coupling term, operating on a

longer spatial scale and whose strength decays with the distance between the coupled population. By including realistic long-range interactions, which, so far, have not been employed in models of spontaneous activity in visual cortex, we also characterized purely spatial properties, such as the lateral spread in cortical space of spontaneous states and the prevalence of mixed states.

The plausibility of the model was further improved by translating the continuous dynamical system into a network of spiking integrate and fire neurons, with the same connectivity structures. In this setting, we explore the idea that a weak modulatory current applied when the network's configuration resembles locally an oriented state can elicit the perception of an elongated oriented feature in a topographically matching location of the visual field. Albeit working in a idealized setting, we conducted a feasibility study which led to establish the relations between stimulation amplitude, temporal resolution and specificity of the percept.

6.2 Extensions

Taken individually, the models presented in the previous chapters, with minimal improvements, can be further exploited to study several other aspects of visual information processing.

6.2.1 Full-size stimuli

We employed the sparse coding model either using a 2-patch configuration, where input field locations where horizontally aligned, or a 5-patch configuration, where direct interactions existed only between the central location and the single surround locations. The former was the minimal setting to present our theory and discuss its potential, while the latter demonstrated that having a bigger surround does not increase the contribution of long-range connections. Of course, neither of the configurations faithfully represent a 'true' visual field.

Whether it is possible to implement a big grid of image patches and learn distance dependent interactions simultaneously remains an interesting question. Learning such a large number of parameters (on the order of tens of millions), often makes neural networks prone to overfitting, a situation when the model is so closely fitted to the training set that it is difficult to generalize and make predictions for new data. To overcome this problem, one could introduce proper regularizations. Regularization is a standard technique in machine learning, that involves adding an extra element to the objective function, which punishes our model for being too complex or, in simple words, for using too high values in the weight matrix, in a similar way that our sparseness constraint does. Other strategies to reduce the number of parameters could consist, for example, in imposing translational symmetries and rotational invariances, or restricting the range of long-range interactions in a way consistent with the anatomy of the cortex.

Setting aside potential numerical difficulties, once a complete set of weights has been obtained, the model could be used to predict how neural responses would look like using full-field stimuli, e.g. pictures of natural scenes where entire object, not only parts of them, are present. Since technology advances in recording technique already allows acquisition of large amount of data simultaneously from many cortical sites, having a model of large cortical patch will be useful to make comparisons (and predictions) with real neural responses.

6.2.2 Higher stages of visual processing

An open question in vision research is how to model the subsequent processing stages that lead to the complete representation of an object from elementary features. Currently, a popular

approach to this question consists in training deep convolutional networks to perform a task, such as object recognition, and then try to compare the emerging interactions to their biological counterparts (Kriegeskorte, 2015; Kindel et al., 2019). Taking inspiration from the hierarchical architecture of the visual streams, deep networks process information through a long sequence of hidden layers (typically 5 to 20), gradually transforming a visual representation, whose spatial layout matches the image, to a semantic representation that enables the recognition of object categories. In the last few years, researchers have trained deep networks whose outputs are remarkably similar to neural representation in early (V1) (Cadena et al., 2019), mid-level (V4) (Bashivan et al., 2019) and high-level (IT) cortical stages of the ventral stream, matching human performance on object-categorization tasks (Yamins et al., 2014). More and more studies focus on improving the structures (Spoerer et al., 2017) or the learning rules (Rotermund and Pawelzik, 2019) of these networks to achieve a closer match with the biology of vision. However, this approach suffers from a fundamental problem: deep networks are very hard to interpret, especially because their architecture, in terms of number of layers and computational units, is chosen a priori and does not really match the anatomy of the visual system. In contrast, our framework offers a way to link the quantities that make up the sensory stimuli (i.e. the variables of the generative model) with the neural activities and the couplings that build a representation of those stimuli (i.e. the neural inference scheme). The goal here would be to include in the generative model increasingly abstract components of natural images while at the same time building a parallel between the emerging network structures and neural circuits in higher cortical areas. The computational principles to derive a realistic dynamics from these more complex models could be similar to the ones presented in Chapter 3 – indeed in extrastriate areas such as V4 or IT neuronal activity is even more sparse.

6.2.3 Spontaneous activity and criticality

Many experimental results support the idea that cortical resting activity has the signature of a system posed near a critical point (Beggs and Plenz, 2003; Pasquale et al., 2008; Petermann et al., 2009; Palva et al., 2013; Shriki et al., 2013; Schuster, 2014), as reflected by scale-free distributions of spike-bursts, usually termed as *neural avalanches*. The emergence and propagation of local activity patterns that we observed in the temporal dynamics presented in Chapter 5 are characterized by cascades of spikes that groups of neighboring neurons fire in close temporal contiguity. I speculate that the size and duration of such spiking patterns are likely to show power law distributions, a property that would make the behavior of our model even more similar to the behavior of cortical networks. More specifically, it might be possible to show that spontaneous orientation-tuned states preferentially emerge in a regime which is close to a critical state, coinciding with the regions identified by our thorough phase space analysis.

Spontaneous states and visual cortical processing are a major topic of research in the criticality community (e.g. Shew et al. (2015)). However, only a few studies employ realistic encoding scenarios or structured functional connectivity – with few exceptions such as (Tomen and Ernst, 2019). If a link between the cortical state transitions and criticality could be made, then our model would offer a realistic context in which to study the computational advantages that being close to critical dynamics has been hypothesized to offer, such as optimizing stimulus detection or the range of input intensity that the network can process.

6.2.4 Incorporating spontaneous activity into the sparse coding framework

The sparse coding model presented in this thesis, as many other models that are defined from first principles, are learned in an unsupervised way. As such, they can suffer from numerical

artifacts, biases contained in the datasets that are chosen for training, limited size of the networks with which the implementation is carried out or difficulties in tuning regularization parameters. These models should not necessarily be used exactly as they result once learning is achieved. Instead, one should take inspiration from their properties and use them to create new, numerically robust, models. Proceeding along this line of thoughts, I propose that a new model could be built where the structure of the network and the profile of interactions reflect the one obtained through the sparse coding optimization, but a two-dimensional topographical organization of the units is imposed on top of that, for example reflecting the empirical orientation maps obtained in V1. In this way, the network used to study spontaneous activity could be incorporated into the sparse coding framework and the two models could be combined in a unique theory. This would result in a more realistic setting in which the interplay between intrinsically generated activity and its modulation by external input can be investigated, which I believe is at the very core of the mechanisms by which the brain represents the external world.

6.3 Perspectives

In conclusion, the present work provides a better understanding of the cortical mechanisms and the anatomical structures that govern contextual processing and spontaneous activity in primary visual cortex. Moreover, it opened the door for investigating generative models of spatially extended scenes by providing a fundamental theoretical framework which can be straightforwardly extended and be scaled with available computational power (a resource which is constantly increasing). Last, it established a testbed for evaluating more sophisticated electrical stimulation protocols before realizing an actual experiment.

Designing experiments to test the proposed stimulation paradigm would require tight collaboration between theoreticians, electrophysiologists and engineers. First of all, we would need to assess whether it is possible to apply a weak, modulatory current such that only neurons close to their firing threshold will spike. Second, we would need to devise a method to check if oriented percepts actually arise. This could be done, for example, by training an animal to report the presence of a specific oriented stimulus on the screen by pressing a lever or by licking (e.g. a go/no-go paradigm) in the absence of stimulation. Performances on the same task in the presence of the sole train of electrical pulses (no stimulus on the screen) could be compared to assess not only the presence of the artificial oriented phosphene, but also the precision of the elicited angle. Third, to create a stimulation pattern, the encoding of images in terms of oriented edge has also to be tested. A psychophysics study in which human subjects are asked to identify natural objects represented as combination of oriented features could clarify which properties of the encoding (orientation jitter, luminance-contrast, resolution of the sampling points…) are more relevant. If the experiment worked, it would not restore perfect vision in blind patients, but it would certainly improve their interactions with the world, simplifying navigation or locations of objects – perhaps even allowing to grab our symbolic cup of coffee.

A | Extended sparse coding model

In this Appendix, we first outline how to extend the generative model to encode an arbitrary number P of patches and how to formulate it in terms of continuous variables for covering the full visual field and then we briefly report the results obtained performing the contextual-modulation experiments using a different, more general configuration of the visual field.

A.1 Generalization of the model: bigger visual field

The proposed version of the model makes use of a 2-patch visual field, which is the minimal setting to investigate the co-occurrence of features in distant locations of an image. Increasing the size of the visual field is straightforward: it suffices to generalize the expression of the feature representation given by Eq. (3.2) and (3.3) as

$$\boldsymbol{b}^u = \boldsymbol{a}^u + \sum_{v=1}^{P} C^{uv}\boldsymbol{a}^v \tag{A.1}$$

for $u = 1, \ldots, P$, where P represents the total number of patches that make up the visual scene. The corresponding equation for generating the image patch $\boldsymbol{s}^u$ would still read as

$$\boldsymbol{s}^u = \Phi\boldsymbol{b}^u. \tag{A.2}$$

Further generalization of the model is possible, going away from discretized partition of the visual field. Denoting by Ω_a, Ω_b and Ω_s the spatial domains of the representations in $\boldsymbol{a}$ and $\boldsymbol{b}$, and the image, respectively, we could implement a continuous version of the generative model by assuming that a visual image $s(\boldsymbol{r})$, for a particular position $\boldsymbol{r} \in \Omega_S$, is obtained by linear combinations of fundamental features $\phi_i(\boldsymbol{r} - \boldsymbol{r}')$

$$s(\boldsymbol{r}) = \sum_i \int_{\Omega_b} \phi_i(\boldsymbol{r} - \boldsymbol{r}')b_i(\boldsymbol{r}')d\boldsymbol{r}' = \int_{\Omega_b} \Phi(\boldsymbol{r} - \boldsymbol{r}')\boldsymbol{b}(\boldsymbol{r}')d\boldsymbol{r}'. \tag{A.3}$$

where

$$\boldsymbol{b}(\boldsymbol{r}) = \int_{\Omega_a} C(\boldsymbol{r} - \boldsymbol{r}')\boldsymbol{a}(\boldsymbol{r}')\, d\boldsymbol{r}'\,. \tag{A.4}$$

To match the hypothesis made in the main text, one should finally assume the following:

a) Features ϕ_i are localized in visual space and do not extend over a maximum range $r_{\max}$ (Fig. A.1, indicated by extension of the yellow cones)
b) C is intended to capture long-range spatial dependencies extending beyond the range $r_{\max}$ of elementary features (Fig. A.1, indicated by the extension of the black and red cones).

More precisely, to obtain again the discrete formulation presented in Chapter 3, one should take the following steps:

- Assume a 1:1 topographic mapping thus having $\Omega_a = \Omega_b = \Omega_s$, since we interpret Ω_s as the visual field and we link $\boldsymbol{a}$ and $\boldsymbol{b}$ to cortical representations.
- Consider two separate, adjacent image patches $\boldsymbol{s}^u$ and $\boldsymbol{s}^v$, whose center locations $\boldsymbol{r}^u$ and $\boldsymbol{r}^v$ fulfill $|\boldsymbol{r}^v - \boldsymbol{r}^u| = r_{\max}$ (this corresponds to the simplest scenario in which C plays a role).
- Set feature correlations at the same location to 1, i.e. $C^{uu} = C^{vv} = \mathbb{I}$.
- Since features separated by a distance $r_{\max}$ are non-overlapping, omit the integration in (A.4) and obtain the following set of equations:

$$\begin{aligned} \boldsymbol{s}^u &= \Phi \boldsymbol{b}^u & \boldsymbol{b}^u &= \boldsymbol{a}^u + C^{uv}\boldsymbol{a}^v \\ \boldsymbol{s}^v &= \Phi \boldsymbol{b}^v & \boldsymbol{b}^v &= \boldsymbol{a}^v + C^{vu}\boldsymbol{a}^u \end{aligned}$$

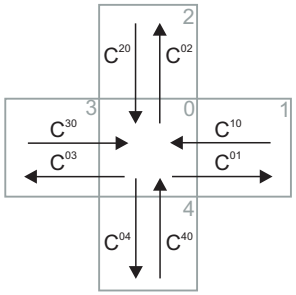
- Assume a *reversal symmetry* $C^{uv}_{ij} = C^{vu}_{ji}$ for all i, j, which implies $C^{vu} = (C^{uv})^\top$.

Remark 1. Equations A.3 and A.4 define the model in a much more general form that what we actually implemented to get the results. Such a formulation allows us to be clear on what are the assumptions we make to derive a simplified patch-model, it shows how one could extend the model to encode bigger image patches and moreover it might be suitable for analytical treatment. However, learning the parameters in such a formulation is not feasible, so we considered a 'patch' version of the model instead.

Remark 2. Figure A.1 illustrates the general philosophy of the generative model we propose. Note that, by including 'objects' at the top of this generative process, we do not mean to suggest a more complicated level of modelling than the one that is described (and used) in Chapter 3 of this thesis. Indeed, as indicated by the dashed lines, we do not explicitly model the numerous operations that lead to the complete representation of an object (e.g. in terms of texture, position, color, border...) from elementary features such as short oriented edges or combinations of edges.

A.2 Contextual modulation with a bigger surround

To check how the responses of the model's units changed when using a 'bigger surround', we repeated the simulations of the three paradigms investigated in the main text using a configuration of the visual field with 4 surround patches instead of only one. Specifically, we consider a visual field composed of five regions, a central patch plus two horizontally and two vertically aligned surround patches, as shown in the image to the right. The same *cross* configuration is assumed for the cortical space, where C^{uv} denotes long-range interactions between distant regions.



For this configuration, the system of differential equations (3.22) easily extends to

$$\begin{cases} \tau_h \dot{\boldsymbol{h}}^u = -\boldsymbol{h}^u + \Phi^\top \boldsymbol{s}^u - \Phi^\top \Phi \boldsymbol{b}^u + \boldsymbol{a}^u \\ \tau_k \dot{\boldsymbol{k}}^u = -\boldsymbol{k}^u + \boldsymbol{a}^u + \sum_{v \in \mathcal{U}(u)} C^{uv}\boldsymbol{a}^v \end{cases}$$

where $\mathcal{U}(u)$ denotes the neighborhood of patch u, with $\mathcal{U}(0) = \{1, 2, 3, 4\}$ and $\mathcal{U}(u) = \{0\}$ for $u = 1, \ldots, 4$. We find that having four surround patches instead of only one does not increase the contribution of long-range connections and does not affect the agreement between our results and experimental data. The surround modulation curves we obtained are shown in Figs. A.2, A.3 and A.4.

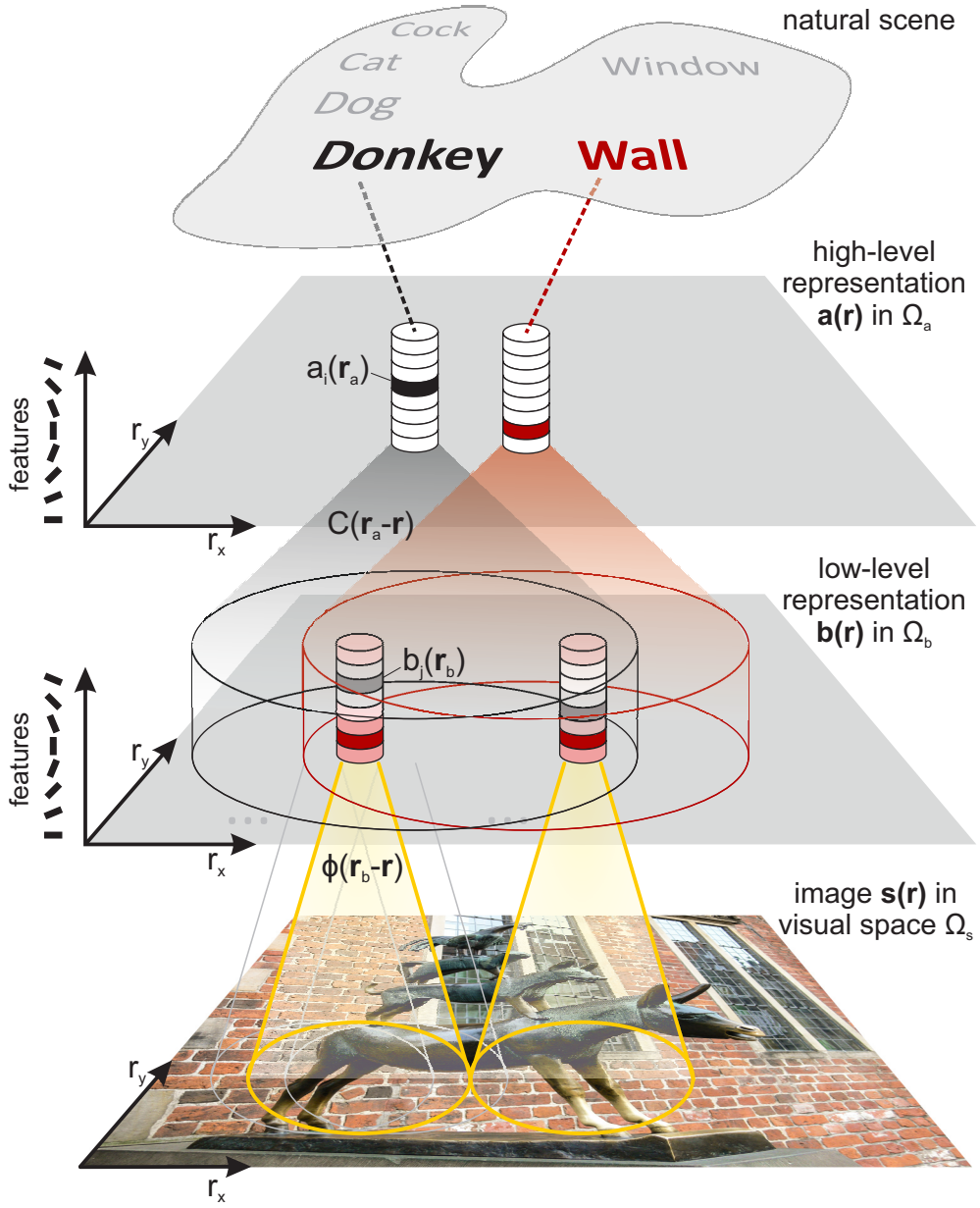


Fig. A.1: Generative model of natural scenes. The content of a visual scene can be described by the objects that compose it – for instance a window, a wall of a building or the statue of a donkey. Such objects imply the presence of particular shapes, textures and contours, such as the vertical legs of the donkey $a_i(r_a)$ or the horizontal bricks of the wall (the dashed lines indicate the numerous processing stages that might lead to the complete representation of an object from elementary features and that we don't model explicitly). These components extend over large regions in the visual field and hence induce long-range spatial correlations $C(r_a - r)$ between more elementary features (black and red shadings and columns). The feature representation $\mathbf{b}$ thus emerges as a superposition of features of both local and more distant image components. The visual image $s(r)$ is finally generated from the feature representation by linear superposition of their (pixel) representations or dictionary Φ.

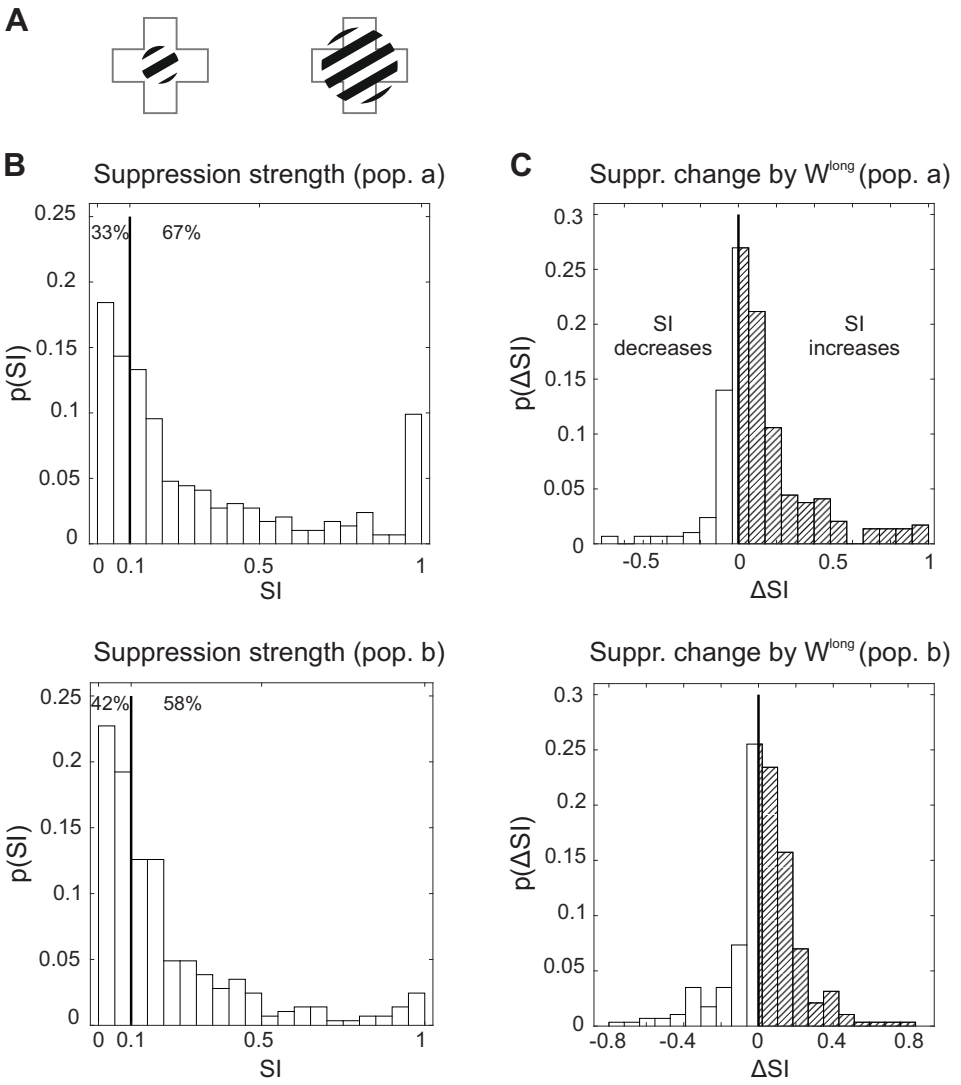


Fig. A.2: Size tuning and surround suppression (4-patch surround). (**A**) Stimulus icons. (**B**) Distribution of suppression indices SI for the full model with long-range interactions. Values of 0 correspond to no suppression, values of 1 to full suppression. (**C**) Change in SI ($\Delta SI = SI^{\text{with long}} - SI^{\text{without}}$) induced by long-range connections. Enhanced suppression occurs more frequently than facilitation in population **a** and, to a lesser extent, in population **b**.

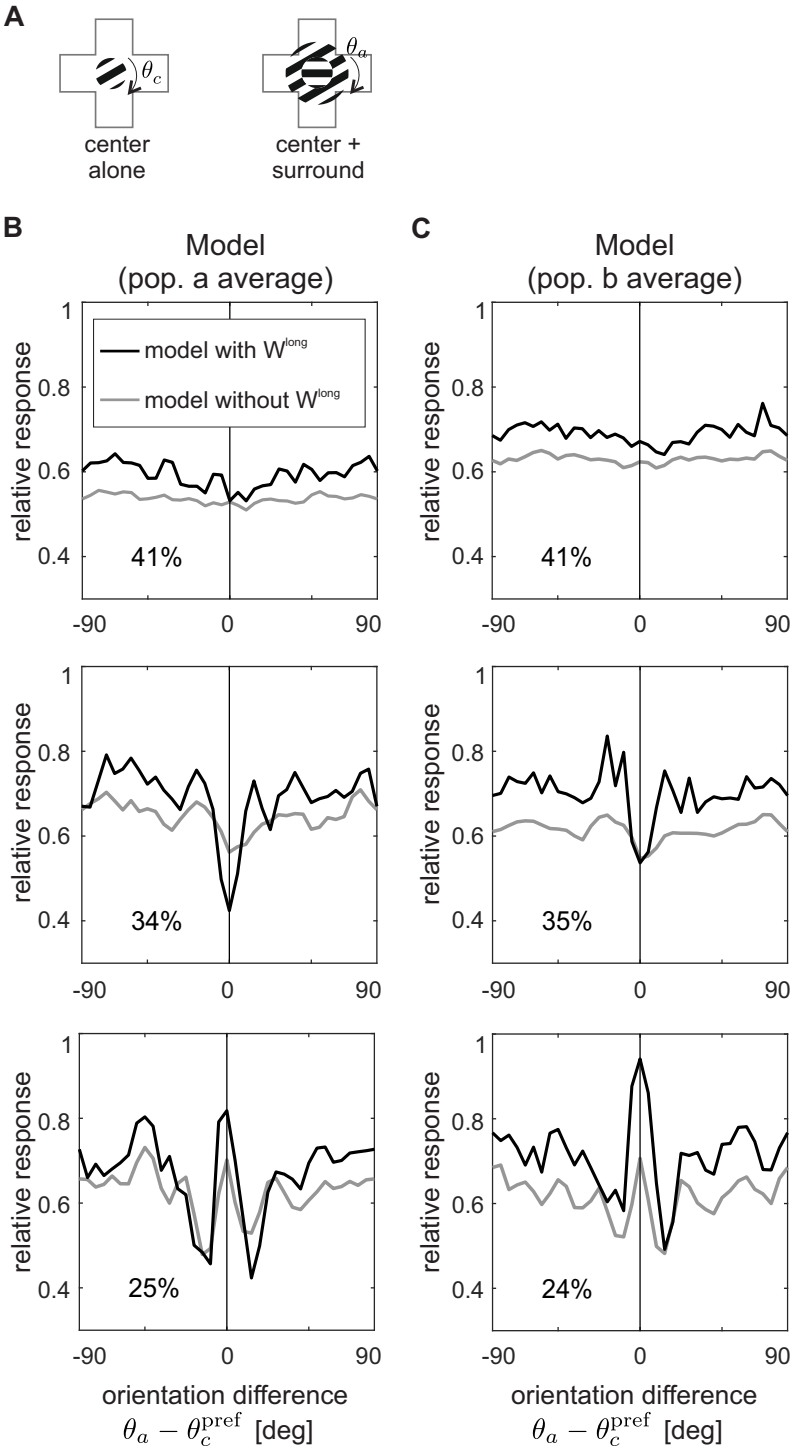


Fig. A.3: Orientation-contrast modulations (4-patch surround). (**A**) Stimulus icons. (**B, C**) Response patterns observed experimentally reproduced by the model (from top to bottom, untuned suppression, iso-orientation suppression and iso-orientation release from suppression) in population **a** and **b** with (black curves) and without (gray curves) long-range interactions to an optimally oriented center stimulus combined with a concentric annulus of varying orientations. Note that responses are shown normalized by the response to the center alone at the preferred orientations of the units. Percentages indicate the proportion of cells that fall in the same orientation-modulation class.

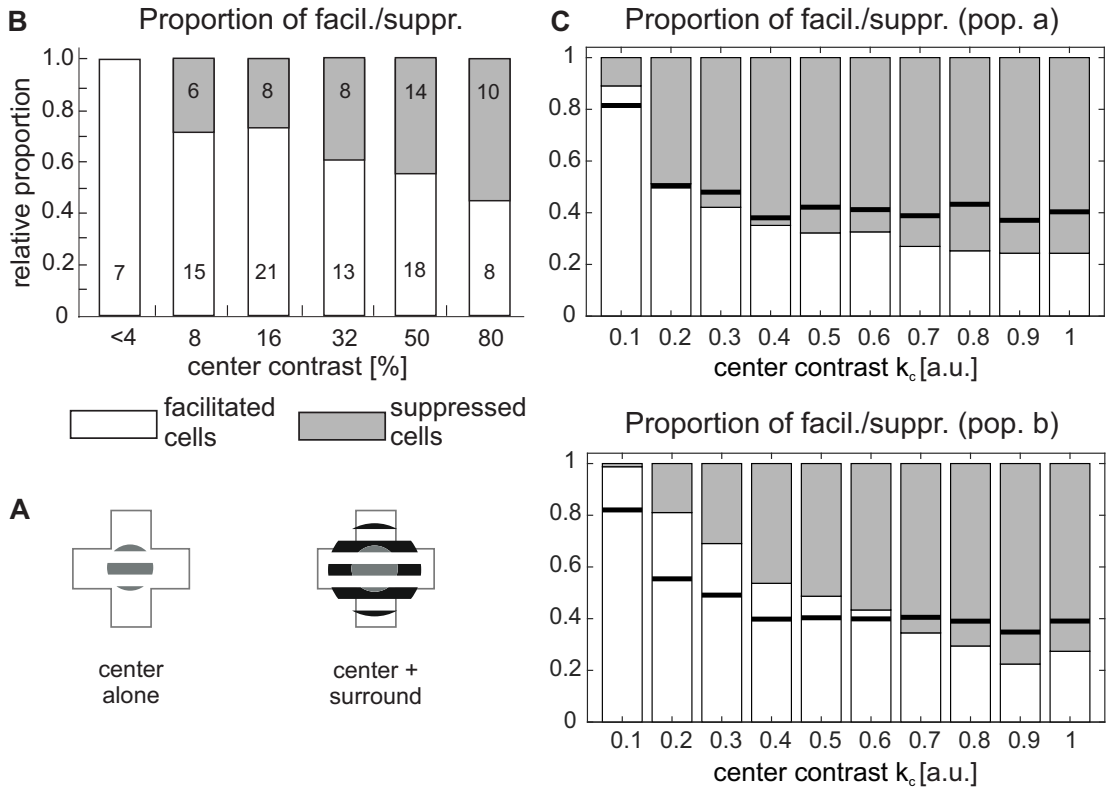


Fig. A.4: Luminance contrast tuning (4-patch surround). (**A**) Stimulus icons. (**B**) Population statistics, detailing the proportion of cells showing facilitation (light bars) or suppression (gray bars) in dependence on center stimulus contrast found in experiments (reproduced from Polat et al. (1998), numbers inside the bars indicate the total number of cells showing suppression/facilitation). (**C**) Population statistics computed from the model's responses of population a (top graph) and b (bottom graph). Cells were judged to be significantly facilitated (suppressed) if their activation ratio between center-surround and center alone stimulation $\mathsf{b}^{\mathrm{sur}}(\mathsf{k_c})/\mathsf{b}^{\mathrm{cen}}(\mathsf{k_c})$ at contrast $\mathsf{k_c}$ was larger than $1+\varepsilon$ (smaller than $1-\varepsilon$), with $\varepsilon = 0.01$. Solid black lines indicate proportion of cells showing facilitation without long-range interactions.

B | Linear stability

In this Appendix, we carry out some calculations useful to analyze the stability of the dynamical system defined in Eq. (4.2), Chapter 4. Specifically, we analyze the eigenvalues of the linearized dynamics $v_\Lambda(\omega)$ (see Eq. (4.11)) and their dependency on the parameter Λ, and we justify the plot shown in Fig. 4.2. Moreover, we carry out the calculations that are need to compute the boundary between the linear and the marginal phase for the case $\Lambda = 0$.

B.1 Eigenvalues spectrum

Consider the family of functions $\Re(v_\Lambda(\omega))$ with

$$v_\Lambda(\omega) = -1 - \gamma J_I \exp\Big(-\frac{\sigma_I^2\omega^2}{2}\Big) + \gamma J_E \frac{1-\Lambda}{1+\Lambda}\sum_{j=-\infty}^{\infty}\Lambda^{|j|}\exp\Big(-\frac{\sigma_E^2\omega^2}{2} + (2\pi i)\omega j\Big). \tag{B.1}$$

For every fixed $\omega > 0$ and every $\Lambda \in (0,1)$, its real part $\Re(v_\Lambda(\omega))$ satisfies the following inequalities

$$\Re(v_1(\omega)) \le \Re(v_\Lambda(\omega)) \le \Re(v_0(\omega)). \tag{B.2}$$

To prove the statement, first we compute a closed-form expression of $\Re(v_\Lambda(\omega))$. Later, we show that it is a decreasing function of Λ, i.e.

$$\Re(v_{\Lambda_1}(\omega)) \ge \Re(v_{\Lambda_2}(\omega)) \text{ for } \Lambda_1 \le \Lambda_2, \tag{B.3}$$

from which the claim follows.

Since the real part of $v_\Lambda(\omega)$ reads as

$$\Re(v_\Lambda(\omega)) = -1 - \gamma J_I \exp\Big(-\frac{\sigma_I^2\omega^2}{2}\Big) + \gamma J_E \exp\Big(-\frac{\sigma_E^2\omega^2}{2}\Big)\frac{1-\Lambda}{1+\Lambda}\sum_{j=-\infty}^{\infty}\Lambda^{|j|}\cos(2\pi j\omega), \tag{B.4}$$

to capture the dependency on Λ, it is convenient to define the family of functions

$$l_\Lambda(\omega) = \frac{1-\Lambda}{1+\Lambda}\sum_{j=-\infty}^{\infty}\Lambda^{|j|}\cos(2\pi j\omega) \tag{B.5}$$

or, equivalently,

$$l_\Lambda(\omega) = \frac{1-\Lambda}{1+\Lambda}\Big[1 + 2\sum_{j=1}^{\infty}\Lambda^{j}\cos(2\pi j\omega)\Big]. \tag{B.6}$$

To find a closed-form expression for $\Re(v_\Lambda(\omega))$, we need to compute the sum in Eq. (B.6). To do this, we observe that the series in Eq. (B.6) is absolutely convergent, since $\Lambda^j|\cos(2\pi j\omega)| \le \Lambda^j$

for all $j \in \mathbb{N}$ and the geometric series $\sum_{j=1}^{\infty} \Lambda^j$ is absolutely convergent. To find its sum, we apply the formula of *summation by parts* twice. Similarly to the integration by parts formula, the summation by part formula states that

$$\sum_{j=m}^{n} a_j \Delta[b_j] = a_{n+1} b_{n+1} - a_m b_m - \sum_{j=m}^{n} \Delta[a_j] b_{j+1}, \tag{B.7}$$

where Δ is the forward difference operator

$$\Delta[a_j] = a_{j+1} - a_j. \tag{B.8}$$

To simplify the notation, we set $z = 2\pi\omega$. Note that

$$\Lambda^j = \Delta\Big[\frac{\Lambda^j}{\Lambda - 1}\Big], \tag{B.9}$$

and, using the trigonometric addition formulas,

$$\begin{aligned}
\Delta[\cos(zj)] &= \cos(z(j+1)) - \cos(zj) = \cos\Big(z\Big(j + \frac{1}{2}\Big) + \frac{z}{2}\Big) - \cos\Big(z\Big(j + \frac{1}{2}\Big) - \frac{z}{2}\Big) \\
&= -2\sin\Big(z\Big(j + \frac{1}{2}\Big)\Big)\sin\Big(\frac{z}{2}\Big), \\
\Delta\Big[\sin\Big(z\Big(j + \frac{1}{2}\Big)\Big)\Big] &= \sin\Big(z(j+1) + \frac{z}{2}\Big) - \sin\Big(z(j+1) - \frac{z}{2}\Big) = 2\cos(z(j+1))\sin\Big(\frac{z}{2}\Big).
\end{aligned}$$

We thus have

$$\begin{aligned}
\sum_{j=1}^{\infty} \Lambda^{|j|}\cos(2\pi j\omega) &= \sum_{j=1}^{\infty} \cos(zj)\Delta\Big[\frac{\Lambda^j}{\Lambda - 1}\Big] = \\
&= \frac{\Lambda^{n+1}}{\Lambda - 1}\cos(z(n+1)) - \frac{\Lambda}{\Lambda - 1}\cos(z) + \frac{2\Lambda}{\Lambda - 1}\sin\Big(\frac{z}{2}\Big)\sum_{j=1}^{n}\sin\Big(z\Big(j + \frac{1}{2}\Big)\Big)\Lambda^j \\
&= \frac{\Lambda^{n+1}}{\Lambda - 1}\cos(z(n+1)) - \frac{\Lambda}{\Lambda - 1}\cos(z) + \frac{2\Lambda}{\Lambda - 1}\sin\Big(\frac{z}{2}\Big)\sum_{j=1}^{n}\sin\Big(z\Big(j + \frac{1}{2}\Big)\Big)\Delta\Big[\frac{\Lambda^j}{\Lambda - 1}\Big] \qquad \text{(B.10)}
\end{aligned}$$

Applying the formula a second time on the last term yields

$$\begin{aligned}
&\sum_{j=1}^{n}\sin\Big(z\Big(j + \frac{1}{2}\Big)\Big)\Delta\Big[\frac{\Lambda^j}{\Lambda - 1}\Big] = \qquad \text{(B.11)} \\
&= \frac{\Lambda^{n+1}}{\Lambda - 1}\sin\Big(z\Big(n + \frac{3}{2}\Big)\Big) - \frac{\Lambda}{\Lambda - 1}\sin\Big(\frac{3}{2}z\Big) - \sum_{j=1}^{n}\Delta\Big[\sin\Big(z\Big(j + \frac{1}{2}\Big)\Big)\Big]\Big(\frac{\Lambda^{j+1}}{\Lambda - 1}\Big) \\
&= \frac{\Lambda^{n+1}}{\Lambda - 1}\sin\Big(z\Big(n + \frac{3}{2}\Big)\Big) - \frac{\Lambda}{\Lambda - 1}\sin\Big(\frac{3}{2}z\Big) - \frac{2}{\Lambda - 1}\sin\Big(\frac{z}{2}\Big)\sum_{j=1}^{n}\cos(z(j+1))\Lambda^{j+1} \qquad \text{(B.12)}
\end{aligned}$$

After observing that

$$\sum_{j=1}^{n}\cos(z(j+1))\Lambda^{j+1} = \sum_{j=1}^{n}\Lambda^j\cos(zj) - \Lambda\cos(z) + \Lambda^{n+1}\cos(z(n+1)),$$

by inserting Eq. (B.12) into Eq. (B.10) and rearranging the terms, we get

$$\begin{aligned}\Big(1+\frac{4\Lambda}{(\Lambda-1)^2}\sin^2\Big(\frac{z}{2}\Big)\Big)\sum_{j=1}^{n}\Lambda^j\cos(z) &= \\ &= -\frac{\Lambda}{\Lambda-1}\cos(z)-\frac{2\Lambda^2}{(\Lambda-1)^2}\sin\Big(\frac{z}{2}\Big)\sin\Big(\frac{3}{2}z\Big)+\frac{4\Lambda^2}{(\Lambda-1)^2}\sin^2\Big(\frac{z}{2}\Big)\cos(z)+\dots \\ &+\Lambda^{n+1}\left[\frac{\cos(z(n+1))}{\Lambda-1}+\frac{2\Lambda}{(\Lambda-1)^2}\sin^2\Big(\frac{z}{2}\Big)\sin\Big(z\big(n+\frac{3}{2}\big)\Big)-\frac{2}{\Lambda-1}\sin\Big(\frac{z}{2}\Big)\cos(z(n+1))\right]\end{aligned} \tag{B.13}$$

The sum is found by taking the limit for $n\to\infty$ (note that the term in square brackets in Eq. (B.13) remain bounded as n grows, while $\Lambda^{n+1}\to 0$), which reads

$$\sum_{j=1}^{\infty}\Lambda^j\cos(z)=\frac{\Lambda\cos(z)-\Lambda^2}{\Lambda^2-2\Lambda\cos(z)+1}. \tag{B.14}$$

Finally, restoring the initial notation, closed expressions for Eq. (B.6) and (B.4) are given by

$$l_\Lambda(\omega)=\frac{(\Lambda-1)^2}{\Lambda^2-2\Lambda\cos(2\pi\omega)+1},\ \text{and} \tag{B.15}$$

$$\Re(\nu_\Lambda(\omega))=-1-\gamma J_I\exp\Big(-\frac{\sigma_I^2\omega^2}{2}\Big)+\gamma J_E\exp\Big(-\frac{\sigma_E^2\omega^2}{2}\Big)\frac{(\Lambda-1)^2}{\Lambda^2-2\Lambda\cos(2\pi\omega)+1}. \tag{B.16}$$

To prove that, for every fixed $\omega>0$, $l_\Lambda(\omega)$ is monotonously decreasing when considered as a function of Λ, we take its first derivative and study its sign, obtaining

$$\frac{d}{d\Lambda}l_\Lambda(\omega)\geq 0\Leftrightarrow 2(\Lambda^2-1)(1-\cos(x))\geq 0\Leftrightarrow(\Lambda^2-1)\geq 0. \tag{B.17}$$

Since we took $0<\Lambda<1$, Eq. (B.17) is never satisfied and l_Λ is thus decreasing.

Finally, note that, as a function of ω, $l_\Lambda(\omega)$ assumes its local optima at integer values of ω (as it is shown in Fig. 4.2). Indeed, local maxima and minima of $l_\Lambda(\omega)$ coincide with maxima and minima of $\cos(2\pi j\omega)$ that are located, respectively, at $\omega=n$ and $\omega=n+\frac{1}{2}$ for $n\in\mathbb{N}$. The correspondent values are

$$l_\Lambda(n)=\frac{1-\Lambda}{1+\Lambda}\sum_{j=-\infty}^{\infty}\Lambda^{|j|}\cos(2nj\pi)=\frac{1-\Lambda}{1+\Lambda}\sum_{j=-\infty}^{\infty}\Lambda^{|j|}=\frac{1-\Lambda}{1+\Lambda}\frac{1+\Lambda}{1-\Lambda}=1$$

$$\begin{aligned}l_\Lambda(n+\frac{1}{2}) &= \frac{1-\Lambda}{1+\Lambda}\sum_{j=-\infty}^{\infty}\Lambda^{|j|}\cos(2nj\pi+j\pi)= \\ &= \frac{1-\Lambda}{1+\Lambda}\Big(1+2\sum_{j=1}^{\infty}\Lambda^j\cos(j\pi)\Big)= \\ &= \frac{1-\Lambda}{1+\Lambda}\Big(1+2\sum_{j\text{ even}}\Lambda^{|j|}-2\sum_{j\text{ odd}}\Lambda^{|j|}\Big)= \\ &= \frac{1-\Lambda}{1+\Lambda}\Big(1+2\sum_{k=1}^{\infty}\Lambda^{2k}-2\sum_{k=0}^{\infty}\Lambda^{2k+1}\Big)=\end{aligned}$$

$$= \frac{1-\Lambda}{1+\Lambda}\Big(1 + 2\sum_{k=1}^{\infty}(\Lambda^2)^k - 2\Lambda\sum_{k=0}^{\infty}(\Lambda^2)^k\Big) =$$
$$= \frac{1-\Lambda}{1+\Lambda}\Big(1 + 2\frac{\Lambda^2}{1-\Lambda^2} - 2\Lambda\frac{1}{1-\Lambda}\Big) =$$
$$= \Big(\frac{1-\Lambda}{1+\Lambda}\Big)^2.$$

B.2 Boundary between linear and marginal phase

Let's define the function

$$f(\omega) := \Re(\nu_0(\omega)) = -1 - \gamma J_I \exp\Big(-\frac{\sigma_I^2\omega^2}{2}\Big) + \gamma J_E \exp\Big(-\frac{\sigma_E^2\omega^2}{2}\Big) \tag{B.18}$$

(see also Eq. (4.14)). Studying its first derivative, we find

$$f'(\omega) \geq 0 \Leftrightarrow \omega^2 \leq (\omega^\star)^2 := \log\Big(\frac{J_I\sigma_I^2}{J_E\sigma_E^2}\Big)\Big(\frac{\sigma_I^2}{2} - \frac{\sigma_E^2}{2}\Big)^{-1} \tag{B.19}$$

If $\frac{J_I\sigma_I^2}{J_E\sigma_E^2} < 1$, then $\omega^\star \notin \mathbb{R}$ and f assumes its maximum at $\omega = 0$. Since $f(0) = -1\gamma J_I + \gamma J_E < 0$, the fixed point is stable. Instead, if $J_I > J_E\frac{\sigma_E^2}{\sigma_I^2}$, then f assumes its maximum at $\omega = \omega^\star$. To find condition for the stability of the fixed point, we need to find when

$$f(\omega^\star) = -\frac{1}{\tau} + \frac{\gamma}{\tau}\Big[J_E\exp\Big(-\frac{(\sigma_E\omega^\star)^2}{2}\Big) - J_I\exp\Big(-\frac{(\sigma_I\omega^\star)^2}{2}\Big)\Big] < 0 \tag{B.20}$$

or, equivalently, when

$$J_E\exp\Big(-\frac{(\sigma_E\omega^\star)^2}{2}\Big) - J_I\exp\Big(-\frac{(\sigma_I\omega^\star)^2}{2}\Big) < \frac{1}{\gamma}. \tag{B.21}$$

We define $S = \sigma_I^2/\sigma_E^2 > 1$, $\Delta S = \sigma_I^2 - \sigma_E^2 > 0$ and plug the expression of $\omega^\star$ into the left hand side of Eq. (B.21), obtaining equivalent conditions

$$J_E\exp\Big(-\frac{\sigma_E^2}{\Delta S}\log\Big(\frac{J_I}{J_E}S\Big)\Big) - J_I\exp\Big(-\frac{\sigma_I^2}{\Delta S}\log\Big(\frac{J_I}{J_E}S\Big)\Big) < \frac{1}{\gamma}$$
$$J_E\Big(\frac{J_I}{J_E}S\Big)^{-\frac{\sigma_E^2}{\Delta S}} - J_I\Big(\frac{J_I}{J_E}S\Big)^{-\frac{\sigma_I^2}{\Delta S}} < \frac{1}{\gamma}$$
$$J_I^{1-\frac{\sigma_I^2}{\Delta S}}J_E^{\frac{\sigma_I^2}{\Delta S}}\Big(S^{-\frac{\sigma_E^2}{\Delta S}} - S^{-\frac{\sigma_I^2}{\Delta S}}\Big) < \frac{1}{\gamma}$$
$$J_I^{-\frac{\sigma_E^2}{\Delta S}}J_E^{\frac{\sigma_I^2}{\Delta S}}S^{-\frac{\sigma_I^2}{\Delta S}}\Big(S-1\Big) < \frac{1}{\gamma}$$
$$J_I^{-\sigma_E^2}J_E^{\sigma_I^2}S^{-\sigma_I^2} < \big(\gamma(S-1)\big)^{-\Delta S}$$
$$J_I < (J_E)^S S^{-S}(\gamma(S-1))^{(S-1)}. \tag{B.22}$$

Equation (B.22) is the analytical expression for the boundary between linear and marginal phase.

Bibliography

Abel, T., Havekes, R., Saletin, J. M., and Walker, M. P. (2013). Sleep, plasticity and memory from molecules to whole-brain networks. *Current Biology*, 23(17):R774–R788.

Amir, Y., Harel, M., and Malach, R. (1993). Cortical hierarchy reflected in the organization of intrinsic connections in macaque monkey visual cortex. *Journal of Comparative Neurology*, 334(1):19–46.

Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., and Bressloff, P. C. (2017). Circuits and mechanisms for surround modulation in visual cortex. *Annual Review of Neuroscience*, 40:425–451.

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J.-M., Bullier, J., and Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22(19):8633–8646.

Angelucci, A. and Shushruth, S. (2013). Beyond the classical receptive field: surround modulation in primary visual cortex. *The New Visual Neurosciences*, pages 425–444.

Arieli, A., Shoham, D., Hildesheim, R., and Grinvald, A. (1995). Coherent spatiotemporal patterns of ongoing activity revealed by real-time optical imaging coupled with single-unit recording in the cat visual cortex. *Journal of Neurophysiology*, 73(5):2072–2093.

Arieli, A., Sterkin, A., Grinvald, A., and Aertsen, A. (1996). Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science*, 273(5283):1868–1871.

Bak, M., Girvin, J., Hambrecht, F., Kufta, C., Loeb, G., and Schmidt, E. (1990). Visual sensations produced by intracortical microstimulation of the human occipital cortex. *Medical and Biological Engineering and Computing*, 28(3):257–259.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436.

Baum, E. B., Moody, J., and Wilczek, F. (1988). Internal representations for associative memory. *Biological Cybernetics*, 59(4-5):217–228.

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

Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

Blasdel, G. G. (1992a). Differential imaging of ocular dominance and orientation selectivity in monkey striate cortex. *Journal of Neuroscience*, 12(8):3115–3138.

Blasdel, G. G. (1992b). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, 12(8):3139–3161.

Blasdel, G. G. and Salama, G. (1986). Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature*, 321(6070):579–585.

Blumenfeld, B., Bibitchkov, D., and Tsodyks, M. (2006). Neural network model of the primary visual cortex: From functional architecture to lateral connectivity and back. *Journal of Computational Neuroscience*, 20(2):219.

Bonhoeffer, T. and Grinvald, A. (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature*, 353(6343):429.

Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of Neuroscience*, 17(6):2112–2127.

Brindley, G. S. and Lewin, W. (1968). The sensations produced by electrical stimulation of the visual cortex. *The Journal of Physiology*, 196(2):479–493.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS Computational Biology*, 15(4):e1006897.

Callaway, E. M. (1998). Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, 21(1):47–74.

Callaway, E. M. (2005). Structure and function of parallel pathways in the primate early visual system. *The Journal of Physiology*, 566(1):13–19.

Casagrande, V. A. and Kaas, J. H. (1994). The afferent, intrinsic, and efferent connections of primary visual cortex in primates. In *Primary visual cortex in primates*, pages 201–259. Springer.

Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, 88(5):2530–2546.

Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of Neurophysiology*, 88(5):2547–2556.

Chao-Yi, L. and Wu, L. (1994). Extensive integration field beyond the classical receptive field of cat's striate cortical neurons—classification and tuning properties. *Vision Research*, 34(18):2337–2355.

Charles, A. S., Yap, H. L., and Rozell, C. J. (2014). Short-term memory capacity in networks via the restricted isometry property. *Neural Computation*, 26(6):1198–1235.

Chen, C.-C., Kasamatsu, T., Polat, U., and Norcia, A. M. (2001). Contrast response characteristics of long-range lateral interactions in cat striate cortex. *Neuroreport*, 12(4):655–661.

Chettih, S. N. and Harvey, C. D. (2019). Single-neuron perturbations reveal feature-specific competition in V1. *Nature*, page 1.

Chiu, C. and Weliky, M. (2001). Spontaneous activity in developing ferret visual cortex in vivo. *Journal of Neuroscience*, 21(22):8906–8914.

Cicmil, N. and Krug, K. (2015). Playing the electric light orchestra–how electrical stimulation of visual cortex elucidates the neural basis of perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1677):20140206.

Coen-Cagli, R., Dayan, P., and Schwartz, O. (2012). Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Computational Biology*, 8(3):e1002405.

Coen-Cagli, R., Kohn, A., and Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, 18(11):1648.

Das, A. and Gilbert, C. D. (1999). Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature*, 399(6737):655.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience*, volume 806. MIT Press, Cambridge, MA.

De Valois, R. L. and De Valois, K. K. (1980). Spatial vision. *Annual Review of Psychology*, 31(1):309–341.

DeAngelis, G. C., Freeman, R. D., and Ohzawa, I. (1994). Length and width tuning of neurons in the cat's primary visual cortex. *Journal of Neurophysiology*, 71(1):347–374.

Desimone, R., Albright, T. D., Gross, C. G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062.

Destexhe, A., Contreras, D., and Steriade, M. (1999). Spatiotemporal analysis of local field potentials and unit discharges in cat cerebral cortex during natural wake and sleep states. *Journal of Neuroscience*, 19(11):4595–4608.

Deuker, L., Olligs, J., Fell, J., Kranz, T. A., Mormann, F., Montag, C., Reuter, M., Elger, C. E., and Axmacher, N. (2013). Memory consolidation by replay of stimulus-specific neural activity. *Journal of Neuroscience*, 33(49):19373–19383.

Dobelle, W. and Mladejovsky, M. (1974). Phosphenes produced by electrical stimulation of human occipital cortex, and their application to the development of a prosthesis for the blind. *The Journal of Physiology*, 243(2):553–576.

Dobelle, W. H. (2000). Artificial vision for the blind by connecting a television camera to the visual cortex. *ASAIO journal*, 46(1):3–9.

Dobelle, W. H., Mladejovsky, M. G., Evans, J. R., Roberts, T., and Girvin, J. (1976). 'Braille' reading by a blind volunteer by visual cortex stimulation. *Nature*, 259(5539):111–112.

Dong, D. W. and Atick, J. J. (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178.

Doty, R. W. (1965). Conditioned reflexes elicited by electrical stimulation of the brain in macaques. *Journal of Neurophysiology*, 28(4):623–640.

Douglas, R. and Martin, K. (1998). Neocortex. In *The Synaptic Organization of the Brain*, pages 459–510. Oxford University Press, 4th edition.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT Press.

Ernst, U., Pawelzik, K., Sahar-Pikielny, C., and Tsodyks, M. (2001). Intracortical origin of visual maps. *Nature Neuroscience*, 4(4):431–436.

Ernst, U. A., Mandon, S., Schinkel-Bielefeld, N., Neitzel, S. D., Kreiter, A. K., and Pawelzik, K. R. (2012). Optimality of human contour integration. *PLoS Computational Biology*, 8(5).

Ernst, U. A., Schiffer, A., Persike, M., and Meinhardt, G. (2016). Contextual interactions in grating plaid configurations are explained by natural image statistics and neural modeling. *Frontiers in Systems Neuroscience*, 10:78.

Felleman, D. J. and Van, D. E. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

Ferezou, I., Bolea, S., and Petersen, C. C. (2006). Visualizing the cortical representation of whisker touch: voltage-sensitive dye imaging in freely moving mice. *Neuron*, 50(4):617–629.

Fitzpatrick, D. (1996). The functional organization of local circuits in visual cortex: insights from the study of tree shrew striate cortex. *Cerebral Cortex*, 6(3):329–341.

Foster, K., Gaska, J. P., Nagler, M., and Pollen, D. (1985). Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *The Journal of Physiology*, 365(1):331–363.

Garrigues, P. and Olshausen, B. A. (2008). Learning horizontal connections in a sparse coding model of natural images. In *Advances in Neural Information Processing Systems*, pages 505–512.

Geisler, W. S., Perry, J. S., Super, B., and Gallogly, D. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724.

Gilbert, C. D. and Wiesel, T. N. (1979). Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. *Nature*, 280(5718):120.

Gilbert, C. D. and Wiesel, T. N. (1983). Clustered intrinsic connections in cat visual cortex. *Journal of Neuroscience*, 3(5):1116–1133.

Gilbert, C. D. and Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, 9(7):2432–2442.

Goldberg, J. A., Rokni, U., and Sompolinsky, H. (2004). Patterns of ongoing activity and the functional architecture of the primary visual cortex. *Neuron*, 42(3):489–500.

Greenberg, D. S., Houweling, A. R., and Kerr, J. N. (2008). Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nature Neuroscience*, 11(7):749.

Grinvald, A. and Hildesheim, R. (2004). VSDI: a new era in functional imaging of cortical dynamics. *Nature Reviews Neuroscience*, 5(11):874.

Haider, B., Krause, M. R., Duque, A., Yu, Y., Touryan, J., Mazer, J. A., and McCormick, D. A. (2010). Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron*, 65(1):107–121.

Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2):400–415.

Hirsch, J. A. and Gilbert, C. D. (1991). Synaptic physiology of horizontal connections in the cat's visual cortex. *Journal of Neuroscience*, 11(6):1800–1809.

Histed, M. H., Bonin, V., and Reid, R. C. (2009). Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. *Neuron*, 63(4):508–522.

Histed, M. H., Ni, A. M., and Maunsell, J. H. (2013). Insights into cortical mechanisms of behavior from microstimulation experiments. *Progress in Neurobiology*, 103:115–130.

Hoyer, P. O. (2002). Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE.

Hu, T., Genkin, A., and Chklovskii, D. B. (2012). A network of spiking neurons for computing sparse representations in an energy-efficient way. *Neural Computation*, 24(11):2852–2872.

Hubel, D. H. (1959). Single unit activity in striate cortex of unrestrained cats. *The Journal of Physiology*, 147(2):226–238.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243.

Hubel, D. H. and Wiesel, T. N. (1974). Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *Journal of Comparative Neurology*, 158(3):295–305.

Hübener, M., Schwarz, C., and Bolz, J. (1990). Morphological types of projection neurons in layer 5 of cat visual cortex. *Journal of Comparative Neurology*, 301(4):655–674.

Hyvärinen, A. and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423.

Hyvärinen, A., Hoyer, P. O., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558.

Iacaruso, M. F., Gasler, I. T., and Hofer, S. B. (2017). Synaptic organization of visual space in primary visual cortex. *Nature*, 547(7664):449.

Issa, N. P., Trepel, C., and Stryker, M. P. (2000). Spatial frequency maps in cat visual cortex. *Journal of Neuroscience*, 20(22):8504–8514.

Iyer, R. and Mihalas, S. (2017). Cortical circuits implement optimal context integration. *bioRxiv*, page 158360.

Jones, B. (1970). Responses of single neurons in cat visual cortex to a simple and a more complex stimulus. *American Journal of Physiology–Legacy Content*, 218(4):1102–1107.

Jones, H., Wang, W., and Sillito, A. (2002). Spatial organization and magnitude of orientation contrast interactions in primate V1. *Journal of Neurophysiology*, 88(5):2796–2808.

Kapadia, M. K., Ito, M., Gilbert, C. D., and Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron*, 15(4):843–856.

Kapadia, M. K., Westheimer, G., and Gilbert, C. D. (2000). Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *Journal of Neurophysiology*, 84(4):2048–2062.

Kara, P. and Boyd, J. D. (2009). A micro-architecture for binocular disparity and ocular dominance in visual cortex. *Nature*, 458(7238):627.

Karklin, Y. and Lewicki, M. S. (2003). Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14(3):483–499.

Karklin, Y. and Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83.

Karlsson, M. P. and Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nature Neuroscience*, 12(7):913.

Kaschube, M. (2014). Neural maps versus salt-and-pepper organization in visual cortex. *Current Opinion in Neurobiology*, 24:95–102.

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., and Arieli, A. (2003). Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–956.

Kindel, W. F., Christensen, E. D., and Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. *Journal of Vision*, 19(4):29–29.

King, P. D., Zylberberg, J., and DeWeese, M. R. (2013). Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. *Journal of Neuroscience*, 33(13):5475–5485.

Knierim, J. J. and Van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of Neurophysiology*, 67(4):961–980.

Ko, H., Hofer, S. B., Pichler, B., Buchanan, K. A., Sjöström, P. J., and Mrsic-Flogel, T. D. (2011). Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87.

Kretzberg, J. Enst, U. (2013). Vision. In Galizia, C. G. and LLedo, P. M., editors, *Neurosciences. From molecule to behavior: a university textbook*. Berlin Heidelberg: Springer.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68.

Landisman, C. E. and Ts'o, D. Y. (2002). Color processing in macaque striate cortex: relationships to ocular dominance, cytochrome oxidase, and orientation. *Journal of Neurophysiology*, 87(6):3126–3137.

Lee, H., Hong, S., Seo, D., Tae, W., and Hong, S. (2000). Mapping of functional organization in human visual cortex: electrical cortical stimulation. *Neurology*, 54(4):849–854.

Levitt, J. B. and Lund, J. S. (1997). Contrast dependence of contextual effects in primate visual cortex. *Nature*, 387(6628):73.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, 12(2):337–365.

Lin, L., Chen, G., Xie, K., Zaia, K. A., Zhang, S., and Tsien, J. Z. (2006). Large-scale neural ensemble recording in the brains of freely behaving mice. *Journal of Neuroscience Methods*, 155(1):28–38.

Lochmann, T., Ernst, U. A., and Deneve, S. (2012). Perceptual inference predicts contextual modulations of sensory responses. *Journal of Neuroscience*, 32(12):4179–4195.

Löwel, S., Bischof, H.-J., Leutenecker, B., and Singer, W. (1988). Topographic relations between ocular dominance and orientation columns in the cat striate cortex. *Experimental Brain Research*, 71(1):33–46.

Lowery, A. J. (2013). Introducing the monash vision group's cortical prosthesis. In *2013 IEEE International Conference on Image Processing*, pages 1536–1539. IEEE.

Lund, J. S. (1973). Organization of neurons in the visual cortex, area 17, of the monkey (macaca mulatta). *Journal of Comparative Neurology*, 147(4):455–495.

Malach, R., Amir, Y., Harel, M., and Grinvald, A. (1993). Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. *Proceedings of the National Academy of Sciences*, 90(22):10469–10473.

McGuire, B. A., Gilbert, C. D., Rivlin, P. K., and Wiesel, T. N. (1991). Targets of horizontal connections in macaque primary visual cortex. *Journal of Comparative Neurology*, 305(3):370–392.

Meister, M., Wong, R. O., Baylor, D. A., and Shatz, C. J. (1991). Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, 252(5008):939–944.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417.

Mizobe, K., Polat, U., Pettet, M. W., and Kasamatsu, T. (2001). Facilitation and suppression of single striate-cell activity by spatially discrete pattern stimuli presented beyond the receptive field. *Visual Neuroscience*, 18(3):377–391.

Murphey, D. K., Maunsell, J. H., Beauchamp, M. S., and Yoshor, D. (2009). Perceiving electrical stimulation of identified human visual areas. *Proceedings of the National Academy of Sciences*, 106(13):5389–5393.

Nauhaus, I., Nielsen, K. J., Disney, A. A., and Callaway, E. M. (2012). Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. *Nature Neuroscience*, 15(12):1683.

Nelson, J. and Frost, B. (1985). Intracortical facilitation among co-oriented, co-axially aligned simple cells in cat striate cortex. *Experimental Brain Research*, 61(1):54–61.

Niven, J. E. and Laughlin, S. B. (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., and Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, 5(11):1203–1209.

O'hashi, K., Fekete, T., Deneux, T., Hildesheim, R., van Leeuwen, C., and Grinvald, A. (2017). Interhemispheric synchrony of spontaneous cortical states at the cortical column level. *Cerebral Cortex*, 28(5):1794–1807.

Ohki, K., Chung, S., Ch'ng, Y. H., Kara, P., and Reid, R. C. (2005). Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, 433(7026):597.

Ohki, K., Chung, S., Kara, P., Hübener, M., Bonhoeffer, T., and Reid, R. C. (2006). Highly ordered arrangement of single neurons in orientation pinwheels. *Nature*, 442(7105):925.

Olmos, A. and Kingdom, F. A. (2004). A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, 33(12):1463–1473.

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(6583):607.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487.

Omer, D. B., Fekete, T., Ulchin, Y., Hildesheim, R., and Grinvald, A. (2018). Dynamic patterns of spontaneous ongoing activity in the visual cortex of anesthetized and awake monkeys are different. *Cerebral Cortex*.

Palva, J. M., Zhigalov, A., Hirvonen, J., Korhonen, O., Linkenkaer-Hansen, K., and Palva, S. (2013). Neuronal long-range temporal correlations and avalanche dynamics are correlated with behavioral scaling laws. *Proceedings of the National Academy of Sciences*, 110(9):3585–3590.

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

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

Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T., and Norcia, A. M. (1998). Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature*, 391(6667):580.

Polat, U. and Sagi, D. (1993). Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vision Research*, 33(7):993–999.

Pollen, D. A. (2004). Brain stimulation and conscious experience. *Consciousness and Cognition*, 13(3):626–645.

Purves, D., Augustine, G., Fitzpatrick, D., Katz, L., LaMantia, A., McNamara, J., Williams, S., et al. (2001). Central visual pathways. *Neuroscience. 2nd Edition. Sunderland, MA: Sinnauer Associates Inc*.

Rao, C. S., Toth, L. J., and Sur, M. (1997). Optically imaged maps of orientation preference in primary visual cortex of cats and ferrets. *Journal of Comparative Neurology*, 387(3):358–370.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87.

Rehn, M. and Sommer, F. T. (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *Journal of Computational Neuroscience*, 22(2):135–146.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88(1):455–463.

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

Ringach, D. L., Shapley, R. M., and Hawken, M. J. (2002). Orientation selectivity in macaque V1: diversity and laminar dependence. *Journal of Neuroscience*, 22(13):5639–5651.

Rockland, K. S. and Lund, J. S. (1983). Intrinsic laminar lattice connections in primate visual cortex. *Journal of Comparative Neurology*, 216(3):303–318.

Rockland, K. S. and Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, 179(1):3–20.

Roelfsema, P. R., Denys, D., and Klink, P. C. (2018). Mind reading and writing: The future of neurotechnology. *Trends in Cognitive Sciences*, 22(7):598–610.

Rotermund, D. and Pawelzik, K. R. (2019). Back-propagation learning in deep spike-by-spike networks. *Frontiers in Computational Neuroscience*, 13:55.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20(10):2526–2563.

Sceniak, M. P., Ringach, D. L., Hawken, M. J., and Shapley, R. (1999). Contrast's effect on spatial summation by macaque V1 neurons. *Nature Neuroscience*, 2(8):733.

Scherf, O., Pawelzik, K., Wolf, F., and Geisel, T. (1999). Theory of ocular dominance pattern formation. *Physical Review E*, 59(6):6977.

Schmidt, E., Bak, M., Hambrecht, F., Kufta, C., O'rourke, D., and Vallabhanath, P. (1996). Feasibility of a visual prosthesis for the blind based on intracortical micro stimulation of the visual cortex. *Brain*, 119(2):507–522.

Schmidt, K. E., Goebel, R., Löwel, S., and Singer, W. (1997). The perceptual grouping criterion of colinearity is reflected by anisotropies of connections in the primary visual cortex. *European Journal of Neuroscience*, 9(5):1083–1089.

Schuster, H. G. (2014). *Criticality in neural systems*. John Wiley & Sons.

Schwabe, L., Obermayer, K., Angelucci, A., and Bressloff, P. C. (2006). The role of feedback in shaping the extra-classical receptive field of cortical neurons: a recurrent network model. *Journal of Neuroscience*, 26(36):9117–9129.

Sengpiel, F., Sen, A., and Blakemore, C. (1997). Characteristics of surround inhibition in cat area 17. *Experimental Brain Research*, 116(2):216–228.

Series, P., Lorenceau, J., and Frégnac, Y. (2003). The "silent" surround of V1 receptive fields: theory and experiments. *Journal of Physiology-Paris*, 97(4-6):453–474.

Shapero, S., Rozell, C., and Hasler, P. (2013). Configurable hardware integrate and fire neurons for sparse approximation. *Neural Networks*, 45:134–143.

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

Shmuel, A., Korman, M., Sterkin, A., Harel, M., Ullman, S., Malach, R., and Grinvald, A. (2005). Retinotopic axis specificity and selective clustering of feedback projections from V2 to V1 in the owl monkey. *Journal of Neuroscience*, 25(8):2117–2131.

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

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., and Davls, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378(6556):492.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216.

Sincich, L. C. and Blasdel, G. G. (2001). Oriented axon projections in primary visual cortex of the monkey. *Journal of Neuroscience*, 21(12):4416–4426.

Smith, G. B., Hein, B., Whitney, D. E., Fitzpatrick, D., and Kaschube, M. (2018). Distributed network interactions and their emergence in developing neocortex. *Nature Neuroscience*, 21(11):1600.

Somers, D. C., Todorov, E. V., Siapas, A. G., Toth, L. J., Kim, D.-S., and Sur, M. (1998). A local circuit approach to understanding integration of long-range inputs in primary visual cortex. *Cerebral Cortex*, 8(3):204–217.

Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in Psychology*, 8:1551.

Stemmler, M., Usher, M., and Niebur, E. (1995). Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. *Science*, 269(5232):1877–1880.

Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press.

Tanaka, Y. and Sagi, D. (1998). Long-lasting, long-range detection facilitation. *Vision Research*, 38(17):2591–2599.

Tehovnik, E. J. and Slocum, W. M. (2013). Electrical induction of vision. *Neuroscience & Biobehavioral Reviews*, 37(5):803–818.

Tomen, N. and Ernst, U. (2019). The role of criticality in flexible visual information processing. In *The Functional Role of Critical Dynamics in Neural Systems*, pages 233–264. Springer.

Toth, L. J., Rao, S. C., Kim, D.-S., Somers, D., and Sur, M. (1996). Subthreshold facilitation and suppression in primary visual cortex revealed by intrinsic signal imaging. *Proceedings of the National Academy of Sciences*, 93(18):9869–9874.

Troyk, P. R. (2017). The intracortical visual prosthesis project. In *Artificial Vision*, pages 203–214. Springer.

Ts'o, D. Y., Gilbert, C. D., and Wiesel, T. N. (1986). Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *Journal of Neuroscience*, 6(4):1160–1170.

Tsodyks, M., Kenet, T., Grinvald, A., and Arieli, A. (1999). Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286(5446):1943–1946.

Valverde, F. (1971). Short axon neuronal subsystems in the visual cortex of the monkey. *International Journal of Neuroscience*, 1(3):181–197.

Van Essen, D. C., Newsome, W. T., and Maunsell, J. H. (1984). The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. *Vision Research*, 24(5):429–448.

Vinje, W. E. and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276.

Von Helmholtz, H. (1860/1962). Handbuch der physiologischen optik. *& Trans. by JPC Southall. Dover English Edition*.

Walker, G. A., Ohzawa, I., and Freeman, R. D. (1999). Asymmetric suppression outside the classical receptive field of the visual cortex. *Journal of Neuroscience*, 19(23):10536–10553.

Walker, G. A., Ohzawa, I., and Freeman, R. D. (2000). Suppression outside the classical cortical receptive field. *Visual Neuroscience*, 17(3):369–379.

Wang, G., Ding, S., and Yunokuchi, K. (2003). Difference in the representation of cardinal and oblique contours in cat visual cortex. *Neuroscience Letters*, 338(1):77–81.

Wang, G., Grone, B., Colas, D., Appelbaum, L., and Mourrain, P. (2011). Synaptic plasticity in sleep: learning, homeostasis and disease. *Trends in Neurosciences*, 34(9):452–463.

Weliky, M., Bosking, W. H., and Fitzpatrick, D. (1996). A systematic map of direction preference in primary visual cortex. *Nature*, 379(6567):725–728.

Weliky, M., Kandler, K., Fitzpatrick, D., and Katz, L. C. (1995). Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns. *Neuron*, 15(3):541–552.

Weliky, M. and Katz, L. C. (1999). Correlational structure of spontaneous neuronal activity in the developing lateral geniculate nucleus in vivo. *Science*, 285(5427):599–604.

Williams, L. R. and Thornber, K. K. (2001). Orientation, scale, and discontinuity as emergent properties of illusory contour shape. *Neural Computation*, 13(8):1683–1711.

Wilson, D. A. (2010). Single-unit activity in piriform cortex during slow-wave state is shaped by recent odor experience. *Journal of Neuroscience*, 30(5):1760–1765.

Wilson, H. R. and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80.

Wolf, F., Pawelzik, K., Scherf, O., Geisel, T., and Löwel, S. (2000). How can squint change the spacing of ocular dominance columns? *Journal of Physiology-Paris*, 94(5-6):525–537.

Wolfe, J., Houweling, A. R., and Brecht, M. (2010). Sparse and powerful cortical spikes. *Current Opinion in Neurobiology*, 20(3):306–312.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Yoshimura, Y., Sato, H., Imamura, K., and Watanabe, Y. (2000). Properties of horizontal and vertical inputs to pyramidal cells in the superficial layers of the cat visual cortex. *Journal of Neuroscience*, 20(5):1931–1940.

Yoshioka, T., Blasdel, G. G., Levitt, J. B., and Lund, J. S. (1996). Relation between patterns of intrinsic lateral connectivity, ocular dominance, and cytochrome oxidase-reactive regions in macaque monkey striate cortex. *Cerebral Cortex*, 6(2):297–310.

Zetzsche, C. and Nuding, U. (2005). Nonlinear and higher-order approaches to the encoding of natural scenes. *Network: Computation in Neural Systems*, 16(2-3):191–221.

Zhu, M. and Rozell, C. J. (2013). Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. *PLoS Computational Biology*, 9(8):e1003191.

Zhu, M. and Rozell, C. J. (2015). Modeling inhibitory interneurons in efficient sensory coding models. *PLoS Computational Biology*, 11(7):e1004353.

Zylberberg, J., Murphy, J. T., and DeWeese, M. R. (2011). A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Computational Biology*, 7(10):e1002250.