Author/s:
Lian, Yanbo

Title:
Learning receptive field properties in a biologically plausible model of primary visual cortex

Date:
2019

Persistent Link:
https://hdl.handle.net/11343/233884

Learning receptive field properties in a biologically plausible model of primary visual cortex

Yanbo Lian
ORCID iD: 0000-0002-8018-9848

October, 2019

A thesis submitted in total fulfillment of the requirements
of the degree of Doctor of Philosophy

Department of Biomedical Engineering
The University of Melbourne
Melbourne, VIC 3051, Australia

Abstract

While the eye is a complex structure, it is just the start of an even more complex series of visual information processing centers in the brain. Around 40% of the human cortex is involved in vision processing. The areas of the brain that process visual information are divided into distinct compartments. The primary visual cortex, area V1, is the first stage of visual processing in the cortex. In general, the visual cortex processes visual information in a hierarchical structure: from thalamus to V1, V1 to V2, and V2 to higher levels. However, many of the connections do not follow these strict hierarchical rules and interconnections between brain areas are very common. In V1, simple and complex cells are two distinct categories of cells. However, how the properties of receptive fields (RFs) for simple and complex cells are learned still remain unclear.

Artificial neural networks are bio-inspired and powerful tools for tasks such as image recognition. Some of the networks can even generate RFs that are very similar to the simple cells in V1. However, there are currently many aspects of neural network models of biological systems that are biologically implausible. We might understand biological vision processing better by incorporating biological constraints in artificial neural networks.

This thesis focuses on V1 and builds biologically plausible models for simple and complex cells by incorporating biological constraints in artificial neural networks. Our results demonstrate that a two-layer model of the visual pathway from the lateral geniculate nucleus to V1 that incorporates biological constraints and efficient coding can account for the emergence of many experimental phenomena of simple cells when the model is trained on natural images. The model demonstrates that efficient coding can be implemented by the V1 simple cells using neural circuits with a simple biologically plausible architecture.

Our model of complex cells, based on the Bienenstock, Cooper and Munro (BCM) rule, demonstrates that properties of RFs of complex cells can be learned using a biologically plausible learning rule. Quantitative comparisons between the model and experimental data are performed. Results show that model complex cells can account for the diversity of complex cells found in experimental studies.

These findings help us to better understand biological vision processing and provide us with insights into the general signal processing principles that the visual cortex employs to process visual information.

Declaration of Authorship

I declare that this thesis is compromised of my original work. All results presented here are the output of research conducted by me.
This thesis is fewer than the maximum word limit of $100,000$ words exclusive of tables, figures and bibliographies.

Signed: Yanbo Lian
Date: January 17, 2020

Acknowledgements

I would like to thank my supervisors, Professor Anthony Burkitt, Professor David Grayden, Dr Tatiana Kameneva, and Dr Hamish Meffin for their guidance, encouragement and support. I am very grateful for all the time they dedicated to my project.

Thank you to Professor Anthony Burkitt for being my principal supervisor and helping me with administrative things. Your rich knowledge was a huge help to my project.

Thank you to Professor David Grayden for teaching me not only professional knowledge in the field of computational neuroscience, but also soft skills, such as project management and communication.

Thank you to Dr Tatiana Kameneva for being supportive and always giving prompt feedback for my work.

Thank you to Dr Hamish Meffin for your deep insights and suggestions about my research that gave me a lot of inspiration.

I would like to thank Dr Ali Almasi for discussions, providing me his experimental data, and helping with programming.

I would like to thank my parents-in-law for giving me the continuing encouragement.

Thank you to my church friends for having fun and doing life together: Joseph, Deborah, Tony, Tracy, Jiahui, Sam, Siqi, Paul, David, Owen, Haitian, Ken, Charles, Abe, Kevin, and many others.

I am also extremely thankful for the support and encouragement of my parents and parents-in-law. I cannot be who I am without you.

To my wife Menghan Li, the most important person in my life, thank you for taking care of me, accepting me as who I am, and walking with me though all the ups and downs.

And I thank God for all of these people. I hope that the skills I have learned will be used for His glory.

Publications

The following chapter of this thesis has been published as a peer-reviewed journal paper:

- **Chapter 2:** Lian Y, Grayden DB, Kameneva T, Meffin H and Burkitt AN. (2019) "Toward a Biologically Plausible Model of LGN-V1 Pathways Based on Efficient Coding". Front. Neural Circuits 13:13. doi: 10.3389/fncir.2019.00013.

Parts of the following chapters of this thesis have been presented at peer-reviewed conferences:

- **Chapter 2:** Lian Y, Grayden DB, Kameneva T, Meffin H and Burkitt AN."A Biologically Plausible Neural Model of Visual Pathways Based on Efficient Coding". 10th Australasian Workshop on Neuro-Engineering and Computational Neuroscience, Brisbane, Australia, December 2017.

- **Chapter 2:** Lian Y, Grayden DB, Kameneva T, Meffin H and Burkitt AN."A Biologically Plausible Neural Model of Visual Pathways Based on Efficient Coding". 27th Annual Computational Neuroscience Meeting (CNS*2018), BMC Neuroscience 2018, 19(Suppl 2):P24, Seattle, USA, July 2018.

- **Chapter 4:** Lian Y, Grayden DB, Kameneva T, Meffin H and Burkitt AN."Learning the receptive field properties of complex cells in V1". 28th Annual Computational Neuroscience Meeting (CNS*2019), Barcelona, Spain, July 2019.

Contents

List of Figures

List of Tables

List of Abbreviations

- **Area TE** Anterior part of the inferior temporal
- **Area TEO** Posterior part of the inferior temporal cortex
- **BCM** Bienenstock, Cooper and Munro
- **CNN** Convolutional neural networks
- **DFT** Discrete Fourier transform
- **FFT** Fast Fourier transform
- **ICA** Independent component analysis
- **ISA** Independent subspace analysis
- **IT** Inferior temporal cortex
- **LGN** Lateral geniculate nucleus
- **MT** Middle temporal
- **NBCM** Normalized BCM
- **NIM** Nonlinear input model
- **NL** Nonlinearity
- **PCA** Principle component analysis
- **RF** Receptive field
- **RGC** Retinal ganglion cell
- **SD** Standard deviation
- **SFA** Slow feature analysis
- **SSC** Spiny stellate cell
- **V1** Primary visual cortex

Chapter 1

Background

The brain is a highly complex structure and can perform sophisticated tasks related to our daily life. However, all the functions of the brain are made possible by neural systems consisting of huge numbers neurons. Inspired by the brain structure, artificial neural networks have been developed to solve complex tasks such as object and speech recognition. In this chapter, some background of biological vision processing and artificial neural networks is provided in order to provide the foundation for the results presented in the following chapters.

1.1 Neural processing in visual pathways

1.1.1 Nervous system

The neuron is the fundamental unit of the nervous system. In a simple description of typical structure in the brain, the four principal parts of a neuron are the soma, dendrites, axon and terminal synaptic boutons. The dendrites receive inputs from other neurons. The soma or cell body processes the received inputs and generates the all-or-none activity called an action potential if the voltage of the membrane depolarises beyond a threshold level. The axon propagates the action potentials (also called spikes) to the terminal synaptic boutons that connect the neuron to inputs of other neurons. The synapse enables a neuron to modify the membrane potential of another neuron via the activity of neurotransmitters; the amount that it affects the receiving (post-synaptic) neuron is called the weight of the synapse. The electrical currents generated by neurotransmitters of a neuron can be either entirely excitatory or inhibitory upon all other post-synaptic neurons, which is known as Dale's law (Strata and Harvey, 1999). Dale's law simply states that the synaptic weight cannot change in polarity. Synaptic weights can be adapted by the pre- and post-synaptic neuronal activity, which is known as synaptic plasticity. The human nervous system consists of about 86 billion interconnected neurons and makes use

of the plasticity of the synapses to adapt the synaptic weights for learning and behavioral tasks (Azevedo et al., 2009).

In order to mathematically model neural behavior, spike-based and rate-based models are typically used. Spike-based neural models, such as the Hodgkin–Huxley model (Hodgkin and Huxley, 1952) and integrate-and-fire model (Abbott, 1999; Burkitt, 2006a,b), describe mechanisms of spike generation. These models are often used when the detailed behavior of the neuron is important or when each spike conveys valuable information. In rate-based models, information about neural behavior is modelled using only the rate at which spikes are generated. In reality, spike-based and rate-based models are both important because they can be used to address different scientific questions.

1.1.2 Visual pathways

The vision signal processing hierarchy of the nervous system can be divided into two stages: information processing in the pre- and post-cortical areas.

Early processing

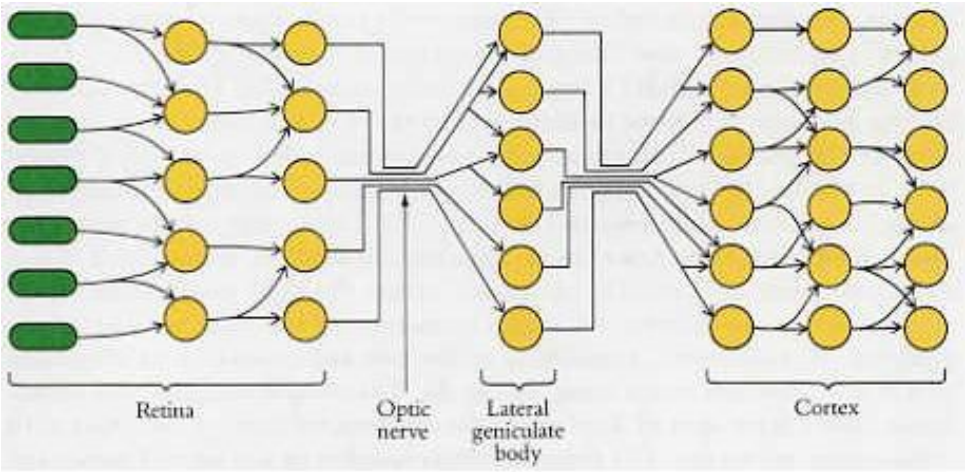

Figure 1.1: The feedforward pathway of early vision processing. Photoreceptor cells and other cells in the retina send information to retinal ganglion cells. Retinal ganglion cells send information via the optic nerve to the lateral geniculate nucleus (also called lateral geniculate body) which is located in the thalamus. The lateral geniculate nucleus then projects to the visual cortex. (Figure from Hubel (1995))

A diagram of early processing pathways is shown in Figure 1.1. Processing starts with the photoreceptor cells of the retina, called rods and cones. When light reaches the photoreceptor cells, light-absorbing visual pigments in the rods and cones cause the cells to hyperpolarize, changing the amount of neurotransmitter released to the next layer, bipolar cells. Rods and cones synapse with bipolar cells whose outputs synapse with

retinal ganglion cells (RGCs) which are the first cells to generate spikes in the visual pathway. There are different classes of RGCs that exhibit different patterns of responses.

RGC axons form the optic nerve. The optic nerve proceeds to the lateral geniculate nucleus (LGN) in the thalamus, which is a deep brain structure. The LGN then sends its outputs to the visual cortex (Hubel, 1995).

In cats, X cells and Y cells, consisting of 70% of the RGC population, are the two well-known groups of RGCs and project to the X and Y cells of LGN (Field and Chichilnisky, 2007). However, Retinal inputs only comprise 5-10% of synapses onto relay neurons in LGN and most of the inputs to LGN are feedback connections from the cortex (Ahmed et al., 1994). The brain uses the huge amount of feedback to maintain stability and to perform many advanced tasks such as attention, recognition, and imagination, but the role of the feedback is still not fully understood (Kruger et al., 2013).

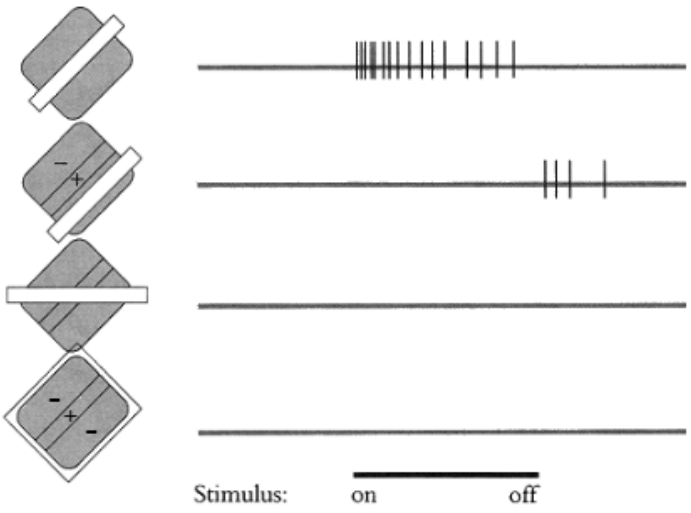

Figure 1.2: The neural response of a simple cell. The stimuli located in the plus (+) region excite responses of the neuron. The stimuli located in the minus (-) region inhibit responses of the neuron. The effective stimuli that cause maximal responses are a solid bar covering the plus (+) region. (Figure adapted from Hubel (1995))

The region of a visual field where a cell is excited and inhibited is called the receptive field (RF) of the cell. The receptive fields (RFs) of retinal ganglion and LGN cells consist of a central zone and a ring-like peripheral surround. The RFs of cortical cells have more complex structures. Hubel and Wiesel (1962) found that cells in cat primary visual cortex (V1) can be divided into simple and complex cells based on their responses to moving bars.

Simple cells are very important in vision processing because they are selective for orientations, spatial frequencies, and spatial phases. The responses of simple cells can be predicted given the stimuli using a linear model, because each simple cell has a small, clearly delineated RF (Hubel, 1995) as shown in Figure 1.2, which shows that the RF of a simple cell has distinguished excitatory and inhibitory regions and the cell has the

maximal response when a bright bar is aligned with the excitatory region of the cell. In addition, Gabor function, a 2-D Gaussian function modulated by a sinusoidal function, can be used to describe the RFs of simple cells (Dayan and Abbott, 2001),

$$D_s(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2})\cos(kx-\phi), \tag{1.1}$$

where σ_x and σ_y determine RF extent in the x and y directions, k is the preferred spatial frequency, and ϕ is the preferred spatial phase.

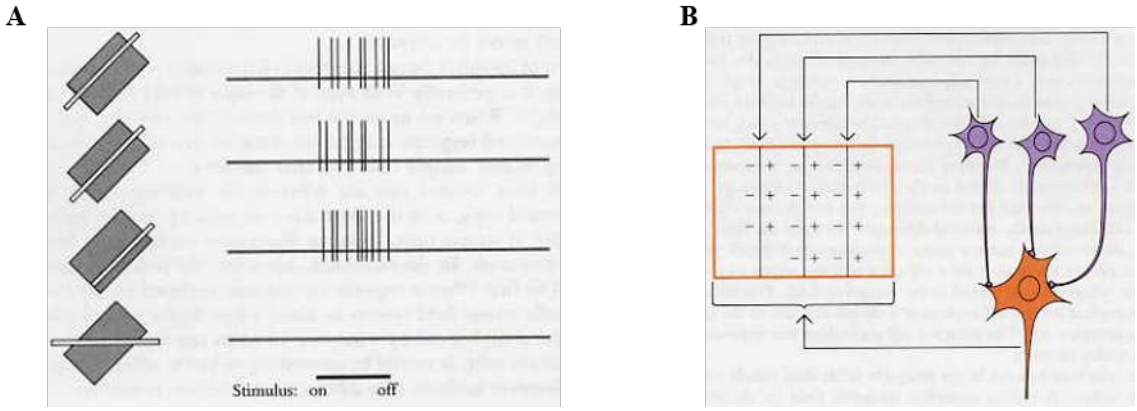

Figure 1.3: The responses and a hierarchical model of a complex cell. This complex cell will respond whenever the oriented stimulus falls within any RF of the three simple cells. (A) Example responses of a complex cell. An oriented bar stimulus evokes a neural response as long as the stimulus is within the RF of the complex cell. (B) Example hierarchical model where three simple cells with spatially different RFs feed into a complex cell. The (-) and (+) regions represent the RF of the cell. (From Hubel (1995))

Complex cells also respond to specifically oriented bars within limited regions of the visual scene similar to simple cells, but their behavior cannot be explained by simply dividing the RFs into excitatory and inhibitory regions (Hubel, 1995). Unlike simple cells, complex cells respond to an appropriately oriented bar at any place within any region of the RF, as shown in Figure 1.3A. This phenomenon can be explained by generating the RF of a complex cell via summation of RFs of multiple simple cells, as shown in Figure 1.3B. Multiple simple cells with RFs that are oriented in the same angle but placed in different locations of the visual scene send information to a complex cell, which causes the complex cell to respond to the stimulus preferred by any of the simple cells from which it receives input. Complex cells are the most common cells in V1 (Hubel and Wiesel, 1968).

Hierarchies in the visual cortex

The visual cortex has different areas of processing visual information. Early visual areas process simple features, while areas deep in the brain process more complex information.

In 1991, Felleman and Essen (1991) proposed a scheme for a comprehensive hierarchy of visual areas.

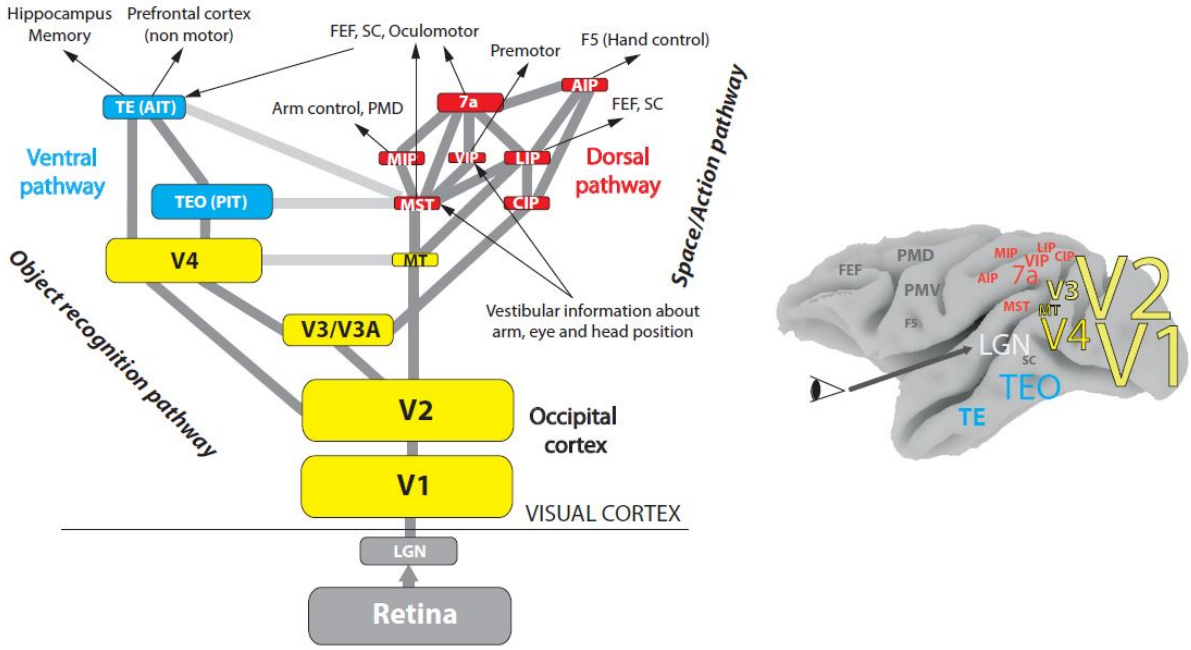

Figure 1.4: Signal pathways and deep hierarchies in visual cortex. The outputs of the retina project to the LGN. The output of the LGN is processed by V1, V2, V4, inferior temporal (area TE) and the posterior part of the inferior temporal cortex (area TEO) along the object recognition (ventral) pathway. The dorsal pathway is involved in action planning and motion processing. (Figure from Kruger et al. (2013))

The hierarchy of visual processing is shown in Figure 1.4. V1 is the first cortical area that processes visual information to generate simple features. Neurons in V2 mostly receive inputs from V1 and respond to more sophisticated contour representations. Neurons in V4 continue integrating lower-level into higher-level responses and become more invariant to size, position or translation of the stimulus. The middle temporal (MT) area is dedicated to visual motion and binocular depth processing. the inferior temporal cortex (IT) is critical for object discrimination and can be partitioned into the anterior part of the inferior temporal (area TE) and the posterior part of the inferior temporal cortex (area TEO) (see Kruger et al. (2013) for a review).

1.1.3 Synaptic transmission in the neocortex

Laminar structure of the neocortex

The neocortex shows a laminar structure and consists of six layers with interneurons in all six layers, spiny stellate cells (SSCs) in layer 4 of primary sensory cortices and pyramidal cells in layers 2-6 (see Silberberg et al. (2005) for a review).

The cortical column, a group of neurons in the cortex within a column of approximately 1 mm^2 area, is usually thought to be the fundamental functional organisation of the cortex (Mountcastle et al., 1955). Figure 1.5 illustrates the layers of a cortical column.

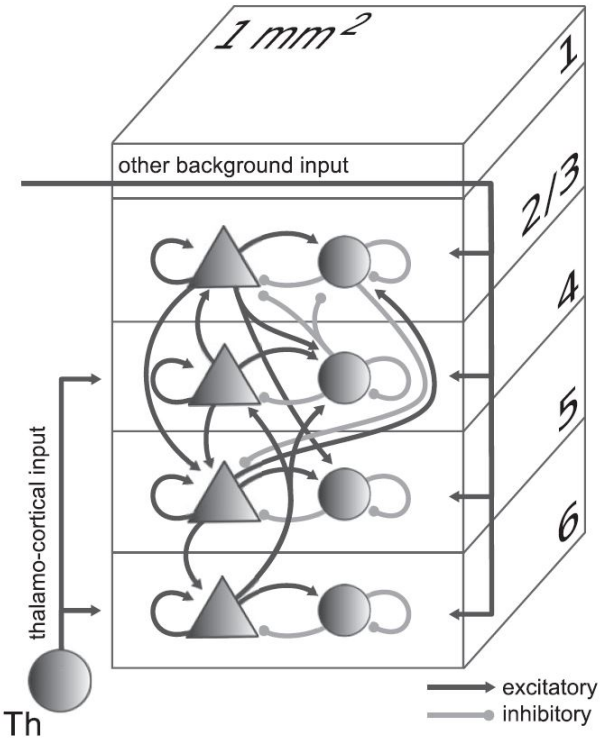

Figure 1.5: The laminar structure of a cortical column. Triangles and circles represent different types of cells (neurons in layer 1 are not shown here). The connectivity in a column consists of intra-laminar (between layers) and inter-laminar (within a layer) connections. (Figure from Potjans and Diesmann (2014))

Four stages of sensory processing

There are four stages of sensory processing. Thalamic input mainly enters into layer 4, which is the first stage, targeting spiny stellate cells (Lund, 1984). Spiny stellate cells send signals to pyramidal neurons in layer 3 (Feldmeyer et al., 2002), which is the second stage of columnar processing. There are strong recurrent connections between pyramidal cells in layer 3 (Silberberg et al., 2005). Layer 2/3 neurons project to layer 5 pyramidal cells, which is the third stage of columnar processing (Thomson and Bannister, 1998). Recurrent connections also exist among layer 5 pyramidal cells (Thomson and Bannister, 2003). The evidence also shows that excitatory connections exist between layer 4 spiny stellate cells and layer 5 pyramidal cells Feldmeyer et al. (2005). The fourth stage of columnar processing is layer 6 pyramidal cells, which receive inputs from layer 5 pyramidal cells (Silberberg et al., 2005).

Layer 6 pyramidal cells, being part of the feedback loop between the cortex and the thalamus, project to the cells in the thalamus (Silberberg et al., 2005). Layer 6 pyramidal

cells largely project to layer 4 (Tarczy-Hornoch et al., 1999) and partly target layer 5 pyramidal neurons (Mercer et al., 2005).

Cell-type specific connectivity

Potjans and Diesmann (2014) used an integrated model to incorporate both anatomical and physiological data to investigate the cell-type specific connections in the cortex. Experimentally derived connection probabilities between different types of cells are shown in Figure 1.6.

Connectivity

		from							
		L2/3e	L2/3i	L4e	L4i	L5e	L5i	L6e	L6i	Th
to	L2/3e	0.101	0.169	0.044	0.082	0.032	0.0	0.008	0.0	0.0
	L2/3i	0.135	0.137	0.032	0.052	0.075	0.0	0.004	0.0	0.0
	L4e	0.008	0.006	0.050	0.135	0.007	0.0003	0.045	0.0	0.0983
	L4i	0.069	0.003	0.079	0.160	0.003	0.0	0.106	0.0	0.0619
	L5e	0.100	0.062	0.051	0.006	0.083	0.373	0.020	0.0	0.0
	L5i	0.055	0.027	0.026	0.002	0.060	0.316	0.009	0.0	0.0
	L6e	0.016	0.007	0.021	0.017	0.057	0.020	0.040	0.225	0.0512
	L6i	0.036	0.001	0.003	0.001	0.028	0.008	0.066	0.144	0.0196

Figure 1.6: The connection probabilities between different types of cells derived by the integrated model of Potjans and Diesmann (2014). There are eight types of cells considered here: excitatory (e) and inhibitory (i) cells in layer 2/3, layer 4, layer 5 and layer 6. The connectivity between thalamus (th) and cortical layers is also included. (Figure from Potjans and Diesmann (2014))

The connectivity probabilities in Figure 1.6 are important because they give a reference for the specific connectivity between any two types of cells in the cortex. This data is very useful for designing a quantitative model of the cortex. More specifically, the thalamus projects mainly to excitatory neurons in layer 4, while most of the inhibitory neurons in layer 4 project to neurons within the same layer. The excitatory neurons in layer 4 mainly synapse with neurons in layer 2/3 and layer 5; there are strong inter-laminar connections in layer 2/3; excitatory neurons in layer 2/3 target excitatory neurons in layer 5. The connection probabilities of these major pathways establish the basis for modelling the microcircuits within the brain.

A canonical microcircuit

A cortical microcircuit that consists of key intrinsic connections is shown in Figure 1.7. This microcircuit includes the classical information processing pathways discussed in the subsection “Four stages of sensory processing” above. However, although this microcircuit description captures our existing knowledge of the cortical circuits, it may not be a complete description of cortical connectivity. As more experimental data on connectivity becomes available, a more comprehensive microcircuit that is consistent with experimental data will emerge that more accurately captures the key functions of the visual cortex.

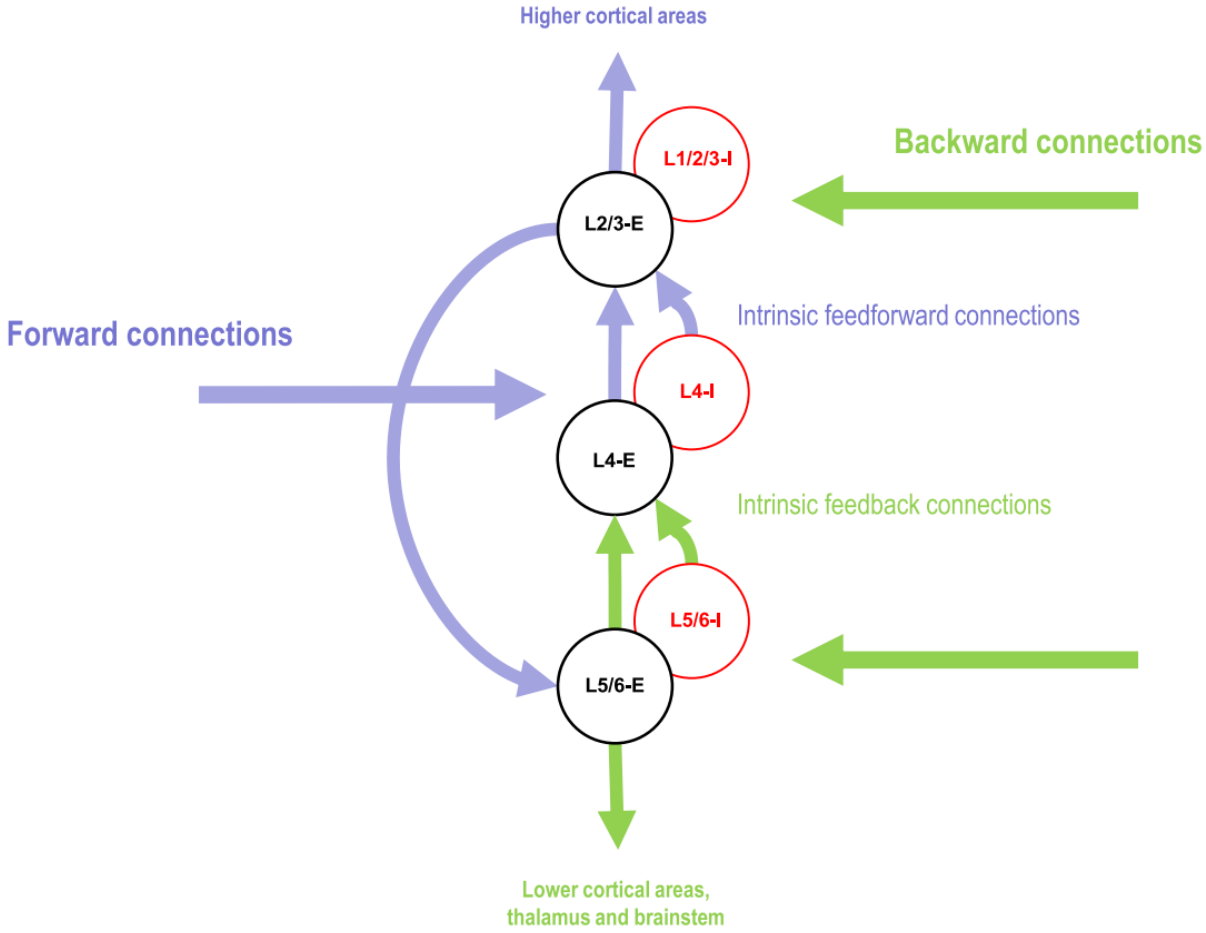

Figure 1.7: The canonical cortical microcircuit. This diagram shows the key intrinsic connections between excitatory (E) and inhibitory (I) populations in layer 2/3, layer 4 and layer 5/6. Afferent input mainly enters layer 4 and is conveyed to layer 2/3. Information from layer 2/3 is then sent to layer 5/6, which sends feedback to layer 4 and earlier cortical areas. There are recurrent connections in each layer. (Figure from Bastos et al. (2012))

1.2 Artificial neural networks

1.2.1 Introduction

The perceptron, conceived in 1957, was the first computational model designed for image recognition (Rosenbaltt, 1957). Even though the perceptron seemed promising at the time, it could not be trained to recognize many classes of patterns, such as the exclusive OR function. However, a multi-layer perceptron network that can achieve nonlinear functions is much more powerful than a single-layer perceptron and can be trained to recognize such classes of patterns. Another famous artificial neural network is the Neocognitron designed by Fukushima (1980). However, the features of the Neocognitron have to be manually designed. Training neural networks that can automatically learn remained challenging for many years until the backpropagation algorithm was applied to learn representations (Rumelhart et al., 1986). The backpropagation procedure employs the calculus chain rule

to compute the derivatives of the output error with respect to each parameter and then makes changes to those parameters in order to reduce the output error. More recently, the development of convolutional neural networks (CNNs) (LeCun and Bengio, 1995) and deep belief networks (Hinton et al., 2006) have had enormous impact on machine learning and artificial intelligence; these networks have been successfully applied to problems such as image and speech recognition and object classification. Cadieu et al. (2014) showed that the performance of a CNN on a core visual object recognition task is comparable to or even outperforms human performance.

Even though the CNN is a biologically-inspired model like other artificial neural networks, it does not account for the biological structure of the visual cortex. In addition, the backpropagation algorithm is purely a mathematical method that is not based upon biological principles of synaptic plasticity. Nevertheless, by studying well-designed artificial neural networks, it is possible that we might gain some insights into how the visual cortex works.

1.2.2 A special case: efficient coding

Efficient coding is a general principle where a neural network represents input stimuli in an efficient way. Two typical models of efficient coding are *sparse coding* and *independent component analysis*. Both models can be implemented by a two-layer neural network, so can be seen as special cases of artificial neural networks. One advantage of efficient coding is that it is energy efficient (Levy and Baxter, 1996). Barlow (1989) claimed that sensory coding aims to reduce the redundancy of statistical dependencies in the inputs.

Sparse coding

Sparse coding, a form of efficient coding, was first proposed by Olshausen and Field (1996). It is known for the ability to develop spatially localized, oriented and bandpass RFs (shown in Figure 1.8) that are similar to the RFs of simple cells in mammalian primary visual cortex (Hubel and Wiesel, 1968; Hawken et al., 1988). The general idea of sparse coding is to find an overcomplete basis set so that the input can be represented by the linear combination of a small number of basis functions in the basis set (Olshausen and Field, 1997). This may save energy for the neural system. The details of sparse coding are presented in Chapter 2.

Independent component analysis

Independent component analysis (ICA) is a linear model that represents data by a linear combination of some basis functions and that makes the coefficients of basis functions as independent as possible (Comon, 1994). ICA often leads to features that characterize the

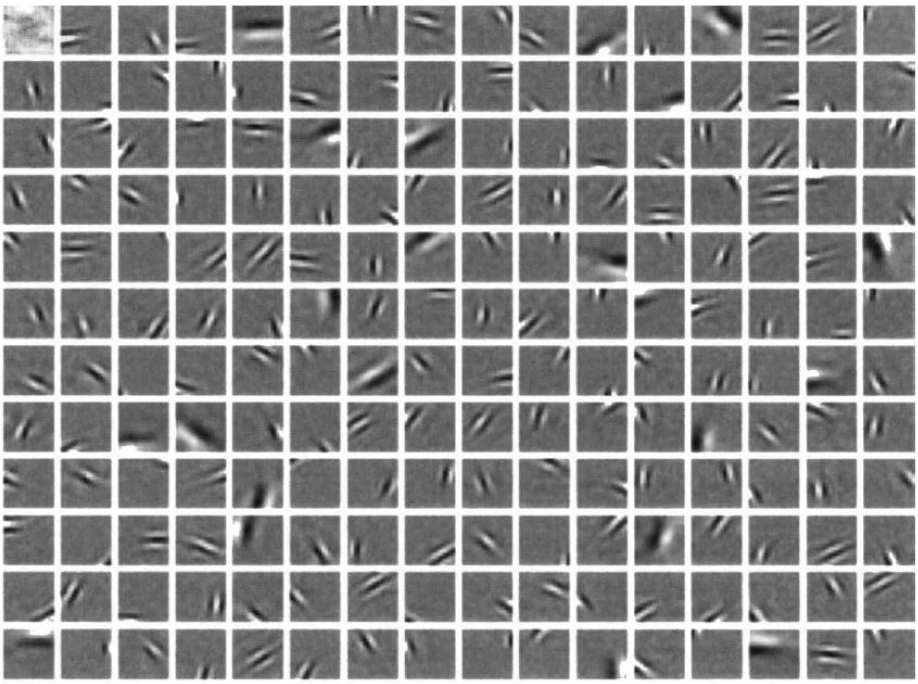

Figure 1.8: Example of sparse coding, where 192 basis images were obtained from 16×16 image patches taken from natural images. The figure is displayed in gray scale (black is negative, white is positive). These basis images are similar to the RFs of simple cells in mammalian primary visual cortex. They are spatially located, oriented and selective to different spatial frequencies. (Figure from Olshausen and Field (1996))

data well. Hateren and Schaaf (1998) applied ICA on natural images and showed that the basis functions were similar to the RFs of simple cells in V1. Therefore, ICA, along with sparse coding, may provide some insights into the formation of V1-like RFs.

1.3 Motivation

Artificial neural networks and the visual cortex have many features in common. First, some networks such as CNNs and deep belief nets, have a hierarchical structure. Second, higher levels of the hierarchy encode increasingly more complex features or advanced functions. Third, feedback plays an important role in both structures, although the role of feedback in the visual system is poorly understood. Fourth, some deep neural networks are able to generate V1-like RFs and to perform object recognition with small error rates.

On the other hand, there are many differences between artificial neural networks and biological neural circuits that implement vision processing. To start with, learning algorithms of artificial neural networks allow the weights to change polarity during the learning process, which is impossible for the biological system. Second, feedback of neural activity plays no role in most artificial neural networks, such as in deep learning, because the feedback is of the error and is not part of the network dynamics. Also, many artificial

neural networks employ the biologically implausible assumption that the feedforward and feedback connections are symmetric. In addition, artificial neural networks cannot directly map onto the microcircuits of the visual cortex because the processing is more complex in the cortex, where recurrent connections always occur in each layer. Furthermore, unlike machine learning algorithms, the role played by feedback in the visual cortex is still unclear.

Therefore, by using bio-inspired neural networks, I aim to discover the mysteries hidden in the visual cortex. How does the visual cortex generate RFs that are useful for further processing? How does the brain achieve complex functions, such as object recognition, in the higher levels of visual processing? What role does the feedback in the visual cortex play? What are the microcircuits the visual cortex employ?

1.4 Outline

In this thesis, biologically plausible models of primary visual cortex are built to explain the RF properties of simple and complex cells. The model is a learning model; i.e. the connection weights connecting different populations are learned when natural images are used as the input stimuli.

In Chapter 2, a biologically plausible model of LGN-V1 pathway is proposed. After the connection weights between LGN and V1 are learned when natural images are used as the input, it is shown that the model can account for experimental phenomena such as ON and OFF sub-region segregation, push-pull effect of simple cells, phase-reversed feedback, diverse RF properties of simple cells and contrast invariance of simple cells. The model is based on efficient coding and uses a variant of sparse coding while incorporating many biological constraints such as Dale's law, local learning rule, and positive neural responses.

In Chapter 3, the principle of efficient coding is investigated to see whether RF properties of complex cells can be learnt when simple cell responses are used as the input to the complex cells. Another layer of complex cells was added onto the simple cell model described in Chapter 2. The learning procedure for LGN-V1 connection is similar to the biologically plausible model of Chapter 2. For learning the connection weights between simple and complex cells, two types of simple cell responses are used: responses of static images and responses of images with temporal information. Results suggest that temporal information is essential for the efficient coding model to pool simple cells into the subspace of complex cells. However, efficient coding suppresses neuronal activities too much, so that model complex cells are highly selective to phase, suggesting that they behave like simple cells.

In Chapter 4, based on my model of the LGN-V1 pathway in Chapter 2, another

layer of complex cells that simply sum simple cell responses in a feedforward fashion was introduced. I apply modified learning rules based on the Bienenstock, Cooper and Munro (BCM) rule (Bienenstock et al., 1982) and show that complex cells can be learned using a biologically plausible mechanism. Furthermore, the resulting model complex cells have a close match to complex cells of cat visual cortex in a recent experimental study, suggesting that the hierarchical structure is important for complex cells to pool simple cells into the subspace. However, the model does not rule out the role of recurrent connections because the model employs responses normalization that might be realized by recurrent activities of neurons.

Finally in Chapter 5, the thesis is concluded with discussion of main findings, some limitations and possible future work.

Chapter 2

Towards a biologically plausible model of LGN-V1 pathways based on efficient coding

Content of this chapter is a slightly modified version of the paper published as *Lian Y, Grayden DB, Kameneva T, Meffin H and Burkitt AN (2019) Toward a Biologically Plausible Model of LGN-V1 Pathways Based on Efficient Coding. Front. Neural Circuits 13:13. doi: 10.3389/fncir.2019.00013* (Lian et al., 2019).

2.1 Abstract

Increasing evidence supports the hypothesis that the visual system employs a sparse code to represent visual stimuli, where information is encoded in an efficient way by a small population of cells that respond to sensory input at a given time. This includes simple cells in primary visual cortex (V1), which are defined by their linear spatial integration of visual stimuli. Various models of sparse coding have been proposed to explain physiological phenomena observed in simple cells. However, these models have usually made the simplifying assumption that inputs to simple cells already incorporate linear spatial summation. This overlooks the fact that these inputs are known to have strong nonlinearities such the separation of ON and OFF pathways, or separation of excitatory and inhibitory neurons. Consequently these models ignore a range of important experimental phenomena that are related to the emergence of linear spatial summation from nonlinear inputs, such as segregation of ON and OFF sub-regions of simple cell receptive fields, the push-pull effect of excitation and inhibition, and phase-reversed cortico-thalamic feedback. Here, we demonstrate that a two-layer model of the visual pathway from the lateral geniculate nucleus to V1 that incorporates these biological constraints on the neural circuits and is based on sparse coding can account for the emergence of

these experimental phenomena, diverse shapes of receptive fields and contrast invariance of orientation tuning of simple cells when the model is trained on natural images. The model suggests that sparse coding can be implemented by the V1 simple cells using neural circuits with a simple biologically plausible architecture.

2.2 Introduction

In early experimental studies of cat striate cortex, Hubel and Wiesel found two main types of cells: simple cells and complex cells (Hubel and Wiesel, 1959, 1962). Simple cells exhibit linear spatial summation of visual stimuli, while complex cells have significant non-linear behavior. This difference is reflected in receptive field (RF) structures of the two types of cells. Receptive fields (RFs) describe spatial patterns of light and dark regions in the visual field that in combination are effective at driving neural response. They are frequently modeled as linear spatial filters. Simple cells have a single RF filter, reflecting the linear spatial summation properties, while complex cells pool the output for two or more RF filters in a nonlinear fashion.

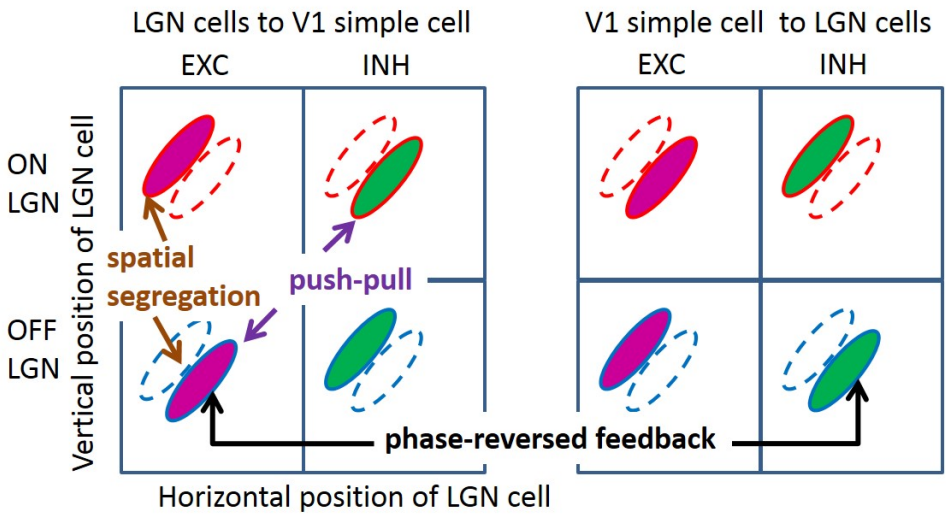

Figure 2.1: Illustration of segregated ON and OFF sub-regions, the push-pull effect, and phase-reversed feedback. ON and OFF LGN cells are spatially located in a 2D region. The colors of magenta and green represent excitatory and inhibitory connections, respectively.

Over the past decades, some important characteristics of simple cell RF have been observed experimentally (with emphasis on cat and primates, but also ferrets). First, simple cells show a range of selectivity for the orientation of visual stimuli, from highly oriented RFs, which are selective to an optimal orientation, to non-oriented RFs, which are insensitive to orientation. Many RFs of simple cells in V1 are oriented, localized, and bandpass (Hubel and Wiesel, 1962, 1968), while non-orientated RFs are seen in all layers of

V1 (Hawken et al., 1988; Chapman and Stryker, 1993). Second, RFs of orientation tuned simple cells can be well described by two-dimensional Gabor functions (Jones and Palmer, 1987a; Ringach, 2002). In addition, both these studies found some blob-like RFs, which are broadly tuned in orientation. Third, RFs of simple cells have spatially segregated ON and OFF sub-regions (Hubel and Wiesel, 1962; Martinez et al., 2005); i.e., the spatial region that excites the simple cell in response to bright (ON) stimuli is separated from the region that excites the cell in response to dark (OFF) stimuli (left column of Figure 2.1). Fourth, simple cells show push-pull responses; i.e., if one stimulus excites a simple cell, the stimulus with opposite contrast, but same location, will inhibit the simple cell (Jones and Palmer, 1987b; Ferster, 1988; Hirsch et al., 1998; Martinez et al., 2005). One example of the push-pull effect can be seen on the left of Figure 2.1 where a simple cell is excited by input from a cell in the lateral geniculate nucleus (LGN) responding to dark spots (an OFF LGN cell) but is effectively inhibited by LGN cells responding a bright spot in the same location (an ON LGN cell). Fifth, feedback from simple cells to LGN cells frequently has a phase-reversed influence compared to the feedforward input (Wang et al., 2006); i.e., where the RF of an ON (OFF) LGN cell is overlapped with the ON (OFF) sub-region of the RF of a simple cell, i.e., feedforward excitation, feedback from the simple cell to the LGN cell is suppressive; where an ON (OFF) LGN cell coincides with the OFF (ON) sub-region of a simple cell RF, i.e., effective feedforward suppression, the feedback is facilitatory. This effect of phase-reversed feedback is also illustrated in Figure 2.1, where the influence from a simple cell to LGN cells is opposite to the influence from LGN cells to the same simple cell. Lastly, the orientation tuning property of simple cells are contrast invariant; i.e., the shape and width of orientation tuning curves remain the same for different stimulus contrasts (Sclar and Freeman, 1982; Skottun et al., 1987; Finn et al., 2007; Priebe, 2016).

On the other hand, insights from computational modelling of V1 cells have also been used to explain experimental data. Sparse coding has been proposed by many researchers as a principle employed by the brain to process sensory information. Olshausen and Field reproduced localized, oriented and spatially bandpass RFs of simple cells based on a *sparse coding* model that aimed to reconstruct the input with minimal average activity of neurons (Olshausen and Field, 1996, 1997). However, the original model failed to generate non-oriented RFs observed in experiments (Ringach, 2002). Subsequently, Olshausen and colleagues found that the sparse coding model can produce RFs that lack strong orientation selectivity by having many more model neurons than the number of input image pixels (Olshausen et al., 2009). Rehn and Sommer introduced *hard sparseness* to classical sparse coding, which minimizes the number of active neurons rather than the average activity of neurons in the original model, and demonstrated that the modified sparse coding model can generate diverse shapes of simple cell RFs (Rehn and Sommer,

2007). Zhu and Rozell showed that many visual non-classical RF effects of V1 such as end-stopping, contrast invariance of orientation tuning can emerge from a dynamical system based on sparse coding (Zhu and Rozell, 2013).

These studies were important in explaining the RF structure, but made a number of simplifying assumptions that overlooked many details of biological reality, include some or all of the following. First, the responses of neurons (e.g., firing rates) should be non-negative. Second, the learning rule of synaptic connections should be local where the changes of synaptic efficacy depend only on pre-synaptic and post-synaptic responses. Third, the learning rule should not violate Dale's Law, namely that neurons release the same type of transmitter at all their synapses, and consequently, the synapses are either all excitatory or all inhibitory (Strata and Harvey, 1999). Fourth, the computation of the response of any neuron should be local, such that only neurons synaptically connected to this target neuron can be involved. In addition, a biologically plausible model should also be consistent with important experimental evidence. For LGN-V1 visual pathways, experimental evidence includes the existence of a large amount of cortico-thalamic feedback (Sherman and Guillery, 1996; Swadlow, 1983), long-range excitatory but not inhibitory connections between LGN and V1, and separated ON and OFF channels for LGN input (Hubel and Wiesel, 1962; Jin et al., 2008, 2011; Ferster et al., 1996). The original sparse coding model neglects many of the biological constraints described above.

Several recent studies addressed the issue of biological plausibility by incorporating some of these constraints, while continuing to neglect others. For example, Zylberberg and colleagues designed a spiking network (based on sparse coding) that can account for diverse shapes of simple cell RFs using lateral inhibition (Zylberberg et al., 2011). The local learning rule and the use of spiking neurons bring some degree of biological plausibility to the model, but the model employs connections that can change sign during learning, which violates Dale's law, and there are not separate channels for ON and OFF LGN input. Additionally, the effect of sparse coding is achieved by competition between units via lateral inhibition, but a recent study suggested that dominant lateral interactions are excitatory in the mouse cortex (Lee et al., 2016). In another modeling work of simple cell RFs, Wiltschut and Hamker designed an efficient coding model with separated ON and OFF LGN cells, and, feedforward, feedback, and lateral connections that can generate various types of simple cell RFs (Wiltschut and Hamker, 2009), but their model does not incorporate Dale's law.

As with earlier studies (Olshausen and Field, 1996, 1997; Rehn and Sommer, 2007; Olshausen et al., 2009), these more recent studies (Wiltschut and Hamker, 2009; Zylberberg et al., 2011), incorporating biological constraints, have continued to focus on the RF structure of simple cells, while largely neglecting the experimental phenomena shown in Figure 2.1. This is because they have typically not separated inputs from ON and OFF

LGN cells, which is a key distinction underlying all the phenomena listed in Figure 2.1. One important question in this regard is how these non-linear (half-wave rectified) LGN inputs are combined to give linear RFs for simple cells and and whether this causes the experimental phenomena listed in Figure 2.1. To our knowledge, Jehee and Ballard are the only researchers that have explicitly explained the effect of phase-reversed feedback using a model based on predictive coding (Jehee and Ballard, 2009). However, the RFs generated by their model do not match well with those observed in experiments and the push-pull effect for simple cells has not been explained. In addition, the formula for calculating responses of model neurons (Jehee and Ballard, 2009, Eq. 7) is not local and the learning rule neglects Dale's law.

In this chapter, we propose a two-layer model of LGN-V1 visual pathways that can account for experimental phenomena:

- segregated ON and OFF sub-regions of simple cells,
- the push-pull effect for simple cells,
- phase-reversed cortico-thalamic feedback,
- diverse shapes of RFs (oriented and non-oriented),
- contrast invariance of orientation tuning.

Our model is biologically plausible by incorporating:

- separate channels of ON and OFF LGN input,
- non-negative neural responses,
- local learning rule,
- Dale's law,
- local computation,
- dynamics of rate-based model neurons,
- feedback from V1 to LGN.

The first layer consists of ON and OFF LGN cells and the second layer consists of simple cells. The connections from the first layer to the second layer (feedforward connections) and from the second layer to the first layer (feedback connections) consist of separate excitatory and inhibitory connections. Even though the inhibitory connections between LGN and V1 should be implemented via intermediate populations of inhibitory interneurons, we use neurons that have both excitatory and inhibitory connections to simplify the circuit.

This aspect of the model is not biologically plausible, but possible biologically plausible neural circuits for implementing inhibitory connections are proposed in the Discussion section. The model presented here is relevant to visual cortices both with and without an orientation columnar organization.

The novelty of the model proposed here is that it models LGN-V1 pathways using segregated ON and OFF LGN channels and separate excitatory and inhibitory connections to investigate the structure of connections between LGN and simple cells to explain a wide range of experimental phenomena. In addition, it can generate a wide variety of experimentally observed RFs of simple cells. Also, the model is biologically plausible by respecting many biological constraints and important experimental evidence. Finally, the experimental phenomena explained in this chapter are all caused by the structure of learned connections between LGN and V1 after the model is trained on natural image data.

2.3 Methods

2.3.1 Sparse coding

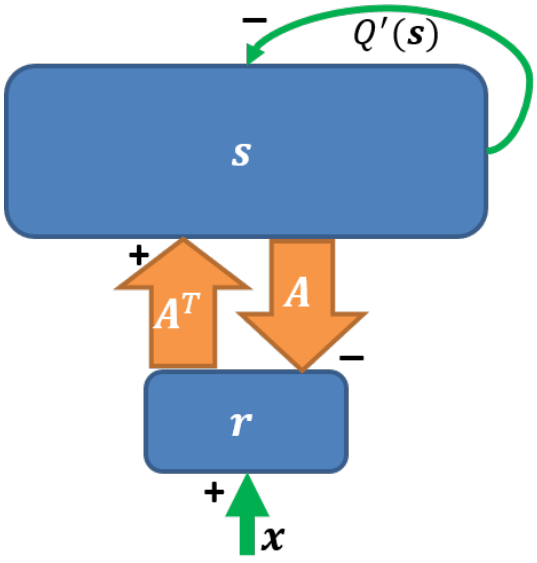

Figure 2.2: The network implementation of sparse coding. Upward and downward arrows represent feedforward and feedback connections. The reconstruction **As** is subtracted via negative feedback. $Q'(\mathbf{s})$ represents self-inhibition of neurons. (Adapted from Figure 5 in (Olshausen and Field, 1997))

The original sparse coding model (Olshausen and Field, 1996) proposed that simple cells represent their sensory input in such a way that their spiking rates in response to

natural images tend to be statistically independent and rarely attain large values (near the top of the cells' dynamic range). Mathematically this means that the joint distribution of spike rates over natural images is the product of the distributions for individual cells, and that each of these individual distributions has a long tail (i.e., high kurtosis). Additionally it was proposed that the representation should allow the reconstruction of the sensory input through a simple weighted sum of visual features with minimal error. This can be formulated as an optimization problem of minimizing the cost function,

$$E(\mathbf{A}, \mathbf{s}) = \frac{1}{2}\|\mathbf{x} - \mathbf{A}\mathbf{s}\|_2^2 + \lambda \sum_i Q(s_i), \tag{2.1}$$

where $\mathbf{x}$ represents the input, columns of the matrix $\mathbf{A}$ represent basis vectors that are universal visual features from which any image can be constructed from a weighted sum, $\mathbf{s}$ is the vector of responses, s_i, of model units that represent the corresponding coefficients for all basis vectors, $Q(\cdot)$ represents a penalty function that favors low activity of model units, and λ is a parameter that scales the penalty function (Olshausen and Field, 1996, 1997). The term $\mathbf{A}\mathbf{s}$ in Eq. 2.1 is the reconstruction of the input from the model, so the first term on the right-hand-side of Eq. 2.1 represents the sum of squared difference between the input and model reconstruction. The second term on the right-hand-side of Eq. 2.1 tends to push $\mathbf{s}$ to small values. Therefore, by solving this minimization problem, the model finds a sparse representation for the input. By taking the partial derivatives of Eq. 2.1 in terms of the elements of $\mathbf{A}$ and $\mathbf{s}$, and applying gradient descent, the dynamic equations and the learning rule are given by

$$\begin{aligned} \dot{\mathbf{s}} &= \mathbf{A}^T\mathbf{r} - \lambda Q'(\mathbf{s}) \\ \Delta\mathbf{A} &\propto \langle \mathbf{r}\mathbf{s}^T \rangle, \end{aligned} \tag{2.2}$$

where $\mathbf{r} = \mathbf{x} - \mathbf{A}\mathbf{s}$, $\langle\cdot\rangle$ is the average operation, the dot notation represents differentiation with regard to time, and $Q'(\cdot)$ represents the derivative of $Q(\cdot)$.

Based on Eq. 2.2, a network implementation of sparse coding, shown in Figure 2.2, was proposed by Olshausen and Field (1997) who suggested that a feedforward-feedback loop can implement sparse coding. The input to the model was natural images that had been whitened using a filter that resembles the center-surround structure of retinal ganglion RFs. However, the original sparse coding model was not biologically plausible in several aspects, such as the possibility of negative spiking rates and the violation of Dale's law. In addition, the input the the model was not split into separate ON and OFF channels. Finally, this network imposed feedback synaptic connections that were anti-symmetric to the corresponding feedforward connections (i.e., equal but opposite in sign) and it was unclear how such symmetry could be achieved using biologically plausible mechanisms.

2.3.2 Structure of our model

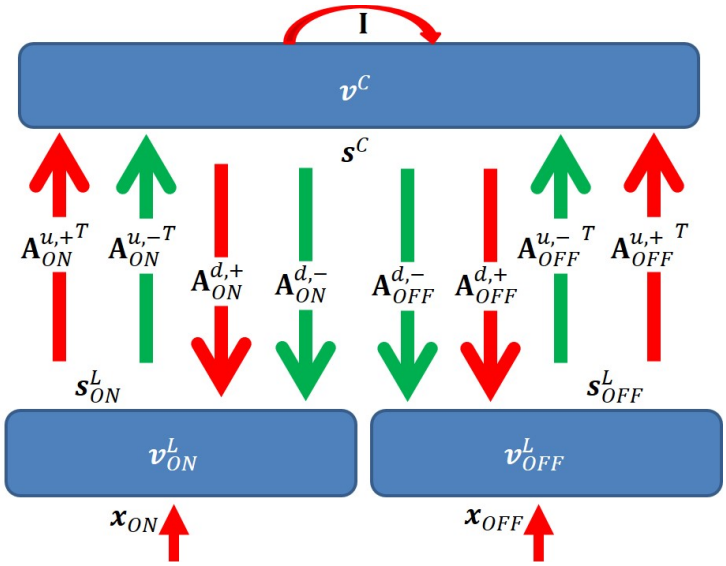

Figure 2.3: Graphical representation of the model. **I** is the identity matrix that represents self-excitation. Red and green arrows represent excitatory and inhibitory connections, respectively. Upward and downward arrows are for feedforward and feedback pathways. Notation defined in the main text.

We propose a two-layer network with rate-based neurons that models the activities of LGN cells (first layer), and simple cells (second layer), respectively (Figure 2.3). The model is based on a locally competitive algorithm that efficiently implements sparse coding with neural dynamics with non-negative spiking rates (Rozell et al., 2008).

We first define the parameters of the model that will be used throughout the chapter. A summary of all symbols defined above is shown in Table 2.1. There are $2N$ LGN cells in the first layer, with N ON LGN cells and N OFF LGN cells, and M simple cells in the second layer. Denote $\mathbf{x} = [x_1, \cdots, x_{2N}]^T$ as the vector of input stimuli to the first layer. Denote $\mathbf{x}_{\text{ON}}$ as the input to ON LGN cells (the first N elements of $\mathbf{x}$) and $\mathbf{x}_{\text{OFF}}$ as the input to OFF LGN cells (the last N elements of $\mathbf{x}$), i.e., $\mathbf{x} = [\mathbf{x}_{\text{ON}}^T, \mathbf{x}_{\text{OFF}}^T]^T$.

Denote $\mathbf{v}^{\text{L}}$ and $\mathbf{s}^{\text{L}}$ as $2N \times 1$ vectors that represent membrane potentials and firing rates of LGN cells in the first layer. Denote $\mathbf{v}^{\text{L}}_{\text{ON}}$, $\mathbf{s}^{\text{L}}_{\text{ON}}$, $\mathbf{v}^{\text{L}}_{\text{OFF}}$, and $\mathbf{s}^{\text{L}}_{\text{OFF}}$ as $N \times 1$ vectors that represent the membrane potentials and firing rates of ON and OFF LGN cells, i.e., $\mathbf{v}^{\text{L}} = [\mathbf{v}^{\text{L}}_{\text{ON}}{}^T, \mathbf{v}^{\text{L}}_{\text{OFF}}{}^T]^T$ and $\mathbf{s}^{\text{L}} = [\mathbf{s}^{\text{L}}_{\text{ON}}{}^T, \mathbf{s}^{\text{L}}_{\text{OFF}}{}^T]^T$. Similarly, $\mathbf{v}^{\text{C}}$ and $\mathbf{s}^{\text{C}}$ are $M \times 1$ vectors that represent membrane potentials and firing rates of M cortical simple cells in the second layer.

In our model, there are several important connections: feedforward (up) excitatory and inhibitory connections from LGN cells to simple cells, feedback (down) excitatory and inhibitory connections from simple cells to LGN cells, and self-excitatory connections of simple cells that represent self-excitation. Definitions of connections are described

Description	**Symbol**
Input stimuli to LGN cells	$\mathbf{x}$
Input stimuli to ON LGN cells	$\mathbf{x}_{\mathrm{ON}}$
Input stimuli to OFF LGN cells	$\mathbf{x}_{\mathrm{OFF}}$
Membrane time constant of LGN cells (12 ms)	τ_{L}
Membrane potentials of LGN cells	$\mathbf{v}^{\mathrm{L}}$
Membrane potentials of ON LGN cells	$\mathbf{v}^{\mathrm{L}}_{\mathrm{ON}}$
Membrane potentials of OFF LGN cells	$\mathbf{v}^{\mathrm{L}}_{\mathrm{OFF}}$
Firing rates of LGN cells	$\mathbf{s}^{\mathrm{L}}$
Firing rates of ON LGN cells	$\mathbf{s}^{\mathrm{L}}_{\mathrm{ON}}$
Firing rates of OFF LGN cells	$\mathbf{s}^{\mathrm{L}}_{\mathrm{OFF}}$
Spontaneous firing rate of LGN cells (2 Hz)	s_{b}
Membrane time constant of LGN cells (12 ms)	τ_{C}
Membrane potentials of cortical simple cells	$\mathbf{v}^{\mathrm{C}}$
Leakage voltages of cortical simple cells	$\mathbf{v}^{\mathrm{C}}_{\mathrm{leak}}$
Firing rates of cortical simple cells	$\mathbf{s}^{\mathrm{C}}$
Excitatory connection: all LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},+}$
Excitatory connection: ON LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},+}_{\mathrm{ON}}$
Excitatory connection: OFF LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},+}_{\mathrm{OFF}}$
Inhibitory connection: all LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},-}$
Inhibitory connection: ON LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},-}_{\mathrm{ON}}$
Inhibitory connection: OFF LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},-}_{\mathrm{OFF}}$
Excitatory connection: simple cells to all LGN cells	$\mathbf{A}^{\mathrm{d},+}$
Excitatory connection: simple cells to ON LGN cells	$\mathbf{A}^{\mathrm{d},+}_{\mathrm{ON}}$
Excitatory connection: simple cells to OFF LGN cells	$\mathbf{A}^{\mathrm{d},+}_{\mathrm{OFF}}$
Inhibitory connection: simple cells to all LGN cells	$\mathbf{A}^{\mathrm{d},-}$
Inhibitory connection: simple cells to ON LGN cells	$\mathbf{A}^{\mathrm{d},-}_{\mathrm{ON}}$
Inhibitory connection: simple cells to OFF LGN cells	$\mathbf{A}^{\mathrm{d},-}_{\mathrm{OFF}}$
Sparsity level (0.6)	λ
Learning rate	η

Table 2.1: Model symbols and parameters in Chapter 2.

below. One aspect of the model that lacks biological plausibility is existence of inhibitory connections between thalamus and cortex, but we propose biologically plausible neural circuits of implementing this aspect of the model in the Discussion section.

Denote $\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},+}$ as an $N \times M$ matrix with non-negative elements that represents the feedforward excitatory connections from ON LGN cells to simple cells. Each column of $\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},+}$ represents connections from N ON LGN cells to a simple cell. Similarly, denote $\mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},+}$ as an $N \times M$ matrix with non-negative elements that represents the feedforward excitatory connections from OFF LGN cells to simple cells. Denote $\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},-}$ and $\mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},-}$ as $N \times M$ matrices with non-positive elements that represent inhibitory connections from ON and OFF LGN cells to simple cells, respectively. Denote $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{u},-}$ as $2N \times M$ matrices that represents all excitatory and inhibitory connections from LGN to V1; then we have $\mathbf{A}^{\mathrm{u},+} = [\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},+}\ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},+}]$ and $\mathbf{A}^{\mathrm{u},-} = [\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},-}\ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},-}]$.

For the feedback pathway, similar notation is used except superscript 'd' represents feedback connections from simple cells to LGN cells. Therefore, we have $\mathbf{A}^{\mathrm{d},+} = [\mathbf{A}_{\mathrm{ON}}^{\mathrm{d},+}\ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},+}]$ and $\mathbf{A}^{\mathrm{d},-} = [\mathbf{A}_{\mathrm{ON}}^{\mathrm{d},-}\ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},-}]$.

Using the notation defined above, the dynamics of ON and OFF LGN cells located in the first layer are given by

$$\begin{aligned} \tau_{\mathrm{L}}\dot{\mathbf{v}}_{\mathrm{ON}}^{\mathrm{L}} &= -\mathbf{v}_{\mathrm{ON}}^{\mathrm{L}} + \mathbf{x}_{\mathrm{ON}} + \mathbf{A}_{\mathrm{ON}}^{\mathrm{d},+}\mathbf{s}^{\mathrm{C}} + \mathbf{A}_{\mathrm{ON}}^{\mathrm{d},-}\mathbf{s}^{\mathrm{C}} + s_{\mathrm{b}} \\ \mathbf{s}_{\mathrm{ON}}^{\mathrm{L}} &= \max(\mathbf{v}_{\mathrm{ON}}^{\mathrm{L}}, 0) \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} \tau_{\mathrm{L}}\dot{\mathbf{v}}_{\mathrm{OFF}}^{\mathrm{L}} &= -\mathbf{v}_{\mathrm{OFF}}^{\mathrm{L}} + \mathbf{x}_{\mathrm{OFF}} + \mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},+}\mathbf{s}^{\mathrm{C}} + \mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},-}\mathbf{s}^{\mathrm{C}} + s_{\mathrm{b}}, \\ \mathbf{s}_{\mathrm{OFF}}^{\mathrm{L}} &= \max(\mathbf{v}_{\mathrm{OFF}}^{\mathrm{L}}, 0), \end{aligned} \tag{2.4}$$

where τ_{L} is the time constant of the membrane potentials of LGN cells, s_{b} is a constant that represents the instantaneous firing rate of the background input (i.e., from neurons outside the network), and the max operation represents the firing dynamics such that a cell only fires when the membrane potential is above a threshold.

Therefore, using the combined notation for ON and OFF LGN cells, the dynamics of LGN cells can be written as

$$\begin{aligned} \tau_{\mathrm{L}}\dot{\mathbf{v}}^{\mathrm{L}} &= -\mathbf{v}^{\mathrm{L}} + \mathbf{x} + (\mathbf{A}^{\mathrm{d},+} + \mathbf{A}^{\mathrm{d},-})\mathbf{s}^{\mathrm{C}} + s_{\mathrm{b}} \\ \mathbf{s}^{\mathrm{L}} &= \max(\mathbf{v}^{\mathrm{L}}, 0). \end{aligned} \tag{2.5}$$

The dynamics of simple cells located in the second layer is given by

$$\begin{aligned} \tau_{\mathrm{C}}\dot{\mathbf{v}}^{\mathrm{C}} = &-(\mathbf{v}^{\mathrm{C}} - \mathbf{v}_{\mathrm{leak}}^{\mathrm{C}}) + \mathbf{A}_{\mathrm{ON}}^{\mathrm{u},+T}\mathbf{s}_{\mathrm{ON}}^{\mathrm{L}} + \mathbf{A}_{\mathrm{ON}}^{\mathrm{u},-T}\mathbf{s}_{\mathrm{ON}}^{\mathrm{L}} \\ &+ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},+\ T}\mathbf{s}_{\mathrm{OFF}}^{\mathrm{L}} + \mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},-\ T}\mathbf{s}_{\mathrm{OFF}}^{\mathrm{L}} + \mathbf{s}^{\mathrm{C}}, \end{aligned} \tag{2.6}$$

which can be reformulated as

$$\begin{aligned} \tau_{\mathrm{C}}\dot{\mathbf{v}}^{\mathrm{C}} &= -\mathbf{v}^{\mathrm{C}} + \mathbf{v}^{\mathrm{C}}_{\mathrm{leak}} + (\mathbf{A}^{\mathrm{u},+} + \mathbf{A}^{\mathrm{u},-})^T\mathbf{s}^{\mathrm{L}} + \mathbf{s}^{\mathrm{C}} \\ \mathbf{s}^{\mathrm{C}} &= \max(\mathbf{v}^{\mathrm{C}} - \lambda, 0), \end{aligned} \quad , \tag{2.7}$$

where τ_{C} is the time constant of the membranes of simple cells and λ is the threshold of the rectifying function of firing rates. In addition, λ is a positive constant that introduces sparseness into the model, $\mathbf{s}^{\mathrm{C}}$ represents the self-excitation of simple cells, which comes from reformulating the model equations of the locally competitive algorithm (Rozell et al., 2008), and $\mathbf{v}^{\mathrm{C}}_{\mathrm{leak}}$ represents the change of membrane potential caused by leakage currents. The leakage currents drive the membrane potentials of simple cells to their resting potentials when there is no external input, i.e., $\mathbf{v}^{\mathrm{C}}$ is zero. Therefore, the steady states of the model dynamics are $\mathbf{v}^{\mathrm{L}} = s_{\mathrm{b}}$, $\mathbf{s}^{\mathrm{L}} = s_{\mathrm{b}}$, $\mathbf{v}^{\mathrm{C}} = 0$, and $\mathbf{s}^{\mathrm{C}} = 0$, which implies that $\mathbf{v}^{\mathrm{C}}_{\mathrm{leak}} = -(\mathbf{A}^{\mathrm{u},+} + \mathbf{A}^{\mathrm{u},-})^T\mathbf{s}_{\mathrm{b}}$, where $\mathbf{s}_{\mathrm{b}}$ is a vector whose elements are all equal to s_{b}. Eqs. 2.5 and 2.7 are solved simultaneously by iteration to obtain values of membrane potentials and firing rates.

2.3.3 Learning rule

The learning process of the model is based on a Hebbian or anti-Hebbian rule, namely that the change of synaptic strength is related only to local pre-synaptic and post-synaptic activities.

The learning rules are given by

$$\begin{aligned} \Delta\mathbf{A}^{\mathrm{u},+} &= \eta\langle(\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}})\mathbf{s}^{\mathrm{C}T}\rangle \\ \Delta\mathbf{A}^{\mathrm{u},-} &= \eta\langle(\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}})\mathbf{s}^{\mathrm{C}T}\rangle \\ \Delta\mathbf{A}^{\mathrm{d},+} &= -\eta\langle(\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}})\mathbf{s}^{\mathrm{C}T}\rangle \\ \Delta\mathbf{A}^{\mathrm{d},-} &= -\eta\langle(\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}})\mathbf{s}^{\mathrm{C}T}\rangle, \end{aligned} \tag{2.8}$$

where η is the learning rate, $\langle\cdot\rangle$ is the ensemble average operation over some samples, $\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}}$ is the vector such that each element of vector $\mathbf{s}^{\mathrm{L}}$ is subtracted by scalar s_{b}, and $(\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}})\mathbf{s}^{\mathrm{C}T}$ is the matrix given by the outer product of vectors $\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}}$ and $\mathbf{s}^{\mathrm{C}}$.

The change of synaptic strength depends only on the pre-synaptic activity ($\mathbf{s}^{\mathrm{L}}$) and post-synaptic activity ($\mathbf{s}^{\mathrm{C}}$). Therefore, this learning rule is local and thus biophysically realistic. In obedience to Dale's law, all the weights of $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{d},+}$ are kept non-negative and all weights of $\mathbf{A}^{\mathrm{u},-}$ and $\mathbf{A}^{\mathrm{d},-}$ are kept non-positive during learning. If any synaptic weight changes sign after applying Eq. 2.8, the synaptic weight is set to zero. In addition, after each learning iteration, synaptic weights are multiplicatively normalized to ensure that Hebbian learning is stable. Specifically, each column of $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{d},-}$

is normalized to norm l_1 and each column of $\mathbf{A}^{\mathrm{u},-}$ and $\mathbf{A}^{\mathrm{d},+}$ is normalized to norm l_2. The multiplicative normalization of synaptic weights may be achieved by homeostatic mechanisms (Turrigiano, 2011), but these are not implemented here as they are not the focus of this chapter.

2.3.4 Input

The data set used in our simulation consists of 10 pre-whitened 512×512 pixel images of natural scenes provided by Olshausen and Field (Olshausen and Field, 1996). Some previous studies of sparse coding (efficient coding) also used this data set (Olshausen and Field, 1996; Zylberberg et al., 2011; Wiltschut and Hamker, 2009; Zhu and Rozell, 2013). The input stimuli to the model are chosen to be 16×16 pixel image patches sampled from these 10 pre-whitened 512×512 pixel images, similar to previous studies (Zylberberg et al., 2011; Zhu and Rozell, 2013).

The pre-whitening process mimics the spatial filtering of retinal processing up to a cut-off frequency determined by the limits of visual acuity (Atick and Redlich, 1992). This process is realized by passing the original natural images through a zero-phase whitening filter with root-mean-square power spectrum,

$$R(f) = f e^{-(f/f_c)^4}, \tag{2.9}$$

where $f_c = 200$ cycles/picture (Olshausen and Field, 1997). Figure 2.4 shows the spatial and frequency profiles of the pre-whitening filter. The spatial profile of the filter (Figure 2.4C), obtained by taking the 2D inverse Fourier transform of the filter in the 2D frequency domain, approximates center-surround RFs of LGN cells in a pixel image. The pre-whitening filter described in Eq. 2.9 is widely used in computational studies (Zhu and Rozell, 2013; Jehee et al., 2006; Jehee and Ballard, 2009; Wiltschut and Hamker, 2009).

The pre-whitened images are then scaled to variance 0.2 similar to Olshausen and Field (1997). Image patches are fed into the first layer, which consists of N ON LGN cells and N OFF LGN cells, i.e., one pixel is fed into one ON LGN cell and one OFF LGN cell. If a pixel intensity in a pre-whitened image patch is negative, we assign the absolute value of the pixel intensity to the input of the OFF LGN cell and set the input of the corresponding ON LGN cell to zero; if the pixel intensity is positive, we set the input of the ON LGN cell to the pixel intensity and set the input to the OFF LGN cell to zero.

2.3.5 Training

Since we use 16×16 pixel images as the input to our model, 256 ON and 256 OFF LGN cells ($N = 256$) are required in the first layer. We use 256 simple cells ($M = 256$) in the second layer. The first-order Euler method is implemented to solve the dynamical system

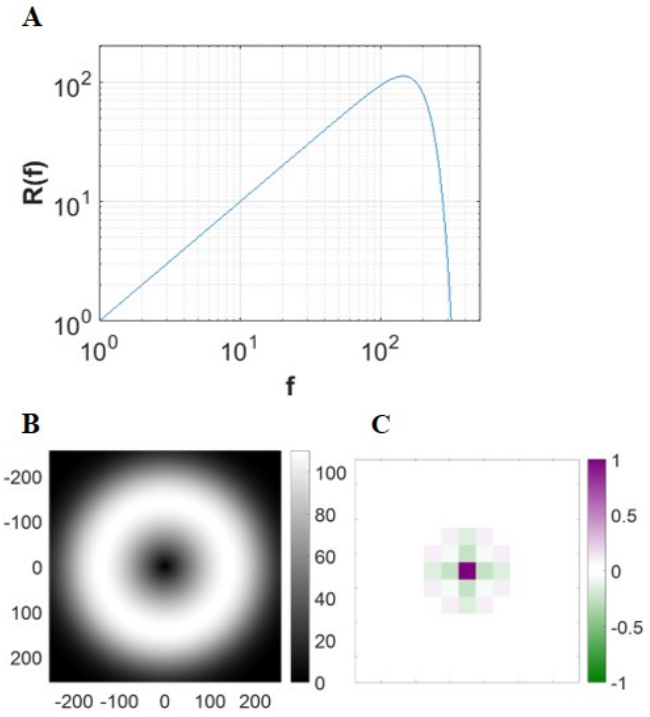

Figure 2.4: Pre-whitening filter. (A) The pre-whitening filter described in Eq. 2.9. (B) The pre-whitening filter in 2D frequency domain. (C) The spatial profile of the pre-whitening filter. The scale of the spatial filter is arbitrarily normalized to convert the luminance to the membrane potential relative to the maximal luminance of the image.

described by Eqs. 2.5 and 2.7. We choose a time scale in which the passive membrane time constant is $\tau_{\mathrm{L}} = \tau_{\mathrm{C}} = 12$ ms, within the physiological range (Dayan and Abbott, 2001), and sparsity level $\lambda = 0.6$ similar to Zhu and Rozell (2013). The spontaneous firing rate, s_{b}, is chosen as $s_{\mathrm{b}} = 2$ Hz, the median of spontaneous firing rates of the mouse LGN cells in the experimental study of Tang et al. (2016). There are 30 integration time steps, with an integration time step of 3ms, for calculating neuronal responses per stimulus with the assumption that neural responses will converge after 30 iterations.

Learning rules in Eq. 2.8 are used to update the synaptic weights. For the normalization step after each learning iteration, each column of $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{d},-}$ is normalized to have norm l_1 and each column of $\mathbf{A}^{\mathrm{u},-}$ and $\mathbf{A}^{\mathrm{d},+}$ is normalized to have norm l_2. Elements of $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{d},+}$ are non-negative and initialized randomly using an exponential distribution with mean 0.5. $\mathbf{A}^{\mathrm{u},-}$ and $\mathbf{A}^{\mathrm{d},-}$ are initialized randomly with non-positive elements that are sampled from an exponential distribution with mean -0.5. Then, synaptic weights are normalized before the learning process starts. Results shown in this chapter are from simulations with $l_1 = l_2 = 1$ (unit norm), as used in the previous study by Rozell et al. (2008). The learning rule based on the average activities of a mini-batch is applied; i.e., in every epoch, a mini-batch that consists of 100 randomly selected 16×16 pixel images sampled from the data set is used. Before the training process of natural image

patches, the model is pre-trained on white noise for 10 000 epochs to mimic the process of pre-development of the visual system; the learning rate is 0.5 in pre-training. To ensure that the weights converge after learning on natural image patches, we use 30 000 epochs in the training process, where the learning rate is 0.5 for the first 10 000 epochs, 0.2 for the second 10 000 epochs and 0.1 for the third 10 000 epochs. Learning rates were chosen to ensure stable convergence of the weights in a reasonable time; but the results are not sensitive to moderate changes. The models described in this chapter are implemented in MATLAB (version R2016b, MathWorks, MA, USA) using my own codes.

2.3.6 Recovering receptive fields of model simple cells using white noise

In order to estimate the RFs of model simple cells in a systematic way, we use the method of spike-triggered averaging to find the pattern that each simple cell responds to on average (Schwartz et al., 2006). Using K 16×16 white noise stimuli $\mathbf{n}_1, \cdots, \mathbf{n}_K$, we present pre-processed stimuli to the model, record the firing rates of a simple cell, $s_1, \cdots, s_K$, and then estimate the RF, $\mathbf{F}$, of the simple cell as the weighted average,

$$\mathbf{F} = \frac{s_1\mathbf{n}_1 + \cdots + s_K\mathbf{n}_K}{s_1 + \cdots + s_K}. \tag{2.10}$$

We used 70 000 white noise stimuli, i.e., $K = 70\ 000$.

In our simulations, we have two versions of estimated RFs using the two different methods of pre-processing the white noise stimuli: the same pre-whitening filter for natural scenes (Eq. 2.9) and a low-pass filter described by

$$L(f) = e^{-(f/f_s)^4}. \tag{2.11}$$

2.3.7 Fitting receptive fields to Gabor functions

The RFs of visual cortical cells are often modelled using a 2D Gabor function $G(x, y)$ of the form

$$\begin{aligned} &G(x, y; x_0, y_0, \sigma_x, \sigma_y, f_s, \beta, \theta, \phi) \\ &\quad = \beta \cos(2\pi f_s x' + \phi) e^{-(\frac{x'}{\sqrt{2}\sigma_x})^2 - (\frac{y'}{\sqrt{2}\sigma_y})^2} \end{aligned} \tag{2.12}$$

with

$$\begin{aligned} x' &= (x - x_0)\cos\theta + (y - y_0)\sin\theta \\ y' &= -(x - x_0)\sin\theta + (y - y_0)\cos\theta, \end{aligned} \tag{2.13}$$

where β is the amplitude, (x_0, y_0) is the center, σ_x and σ_y are standard deviations of the Gaussian envelope, θ is the orientation, f_s is the spatial frequency, and ϕ is the phase of the sinusoid wave (Ringach, 2002). These parameters are fitted using the built-in

MATLAB (version R2016b, MathWorks, MA, USA) function, *lsqcurvefit*, that efficiently solves non-linear curve-fitting problems using a least-squares method. The fitting error is defined as the square of the ratio between the fitting residual and RF.

However, some RFs have large fitting errors, which might give wrong implications when using fitted parameters to represent RFs. In addition, RFs with centres outside the field might also lead to incorrect fitting parameters. Therefore, to ensure that results were only reported for RFs that were well fitted to Gabor functions, we excluded RFs for which either (1) the synaptic fields had fitting error larger than 40% or (2) the center of the fitted Gabor functions lay either outside the block, or within one standard deviation of the Gaussian envelope of the block edge (Zylberberg et al., 2011). After applying these two quality control measures, 140 out of 256 model cells remained for subsequent analysis.

2.3.8 Measuring the overlap index between ON and OFF sub-regions

To investigate the extent of overlap between ON and OFF sub-regions, we used an overlap index that was used in experimental studies (Martinez et al., 2005; Schiller et al., 1976a). Similar to the method used in (Martinez et al., 2005), each ON and OFF excitatory sub-region was fitted by an elliptical Gaussian function:

$$h(x, y; x_0, y_0, a, b, \theta, \gamma) = \frac{\gamma}{2\pi ab} e^{-\frac{x'^2}{2a^2} - \frac{y'^2}{2b^2}} \tag{2.14}$$

where γ is the amplitude, a and b are half axes of the ellipse, and x' and y' are the transformed coordinates given by Eq. 2.13. If there are more than one ON (or OFF) sub-regions for the simple cell, only the most significant sub-region was fitted by the elliptic Gaussian. If either the ON or OFF sub-region of a simple cell has fitting error larger than 40% or has the half axis, a, larger than 3 pixels, this simple cell is excluded. 92 simple cells remained for the analysis of overlap index.

The overlap index, I_{o}, is then defined as

$$I_{\mathrm{o}} = \frac{W_{\mathrm{ON}} + W_{\mathrm{OFF}} - d}{W_{\mathrm{ON}} + W_{\mathrm{OFF}} + d}, \ (-1 < I_{\mathrm{o}} \leq 1) \tag{2.15}$$

where W_{ON} and W_{OFF} are the half width measured at the point where the response is 30% of the maximal response, and d is the distance between the centers of ON and OFF sub-regions. Smaller values of I_{o} indicate more segregation between ON and OFF sub-regions.

2.3.9 Measuring the push-pull index

The push-pull effect of the model was measured by a push-pull index (Martinez et al., 2005). First, for each simple cell, we recorded the membrane potential, P, when the preferred input (the synaptic field) was presented to the model. Next, we recorded the membrane potential, N, while presenting the opposite of preferred input to the model. To make the measurement independent of the relative strength of different simple cells, P and N were normalized by

$$P = \frac{P}{\max(|P|, |N|)} \text{ and } N = \frac{N}{\max(|P|, |N|)}. \tag{2.16}$$

The Push-pull index, I_{p}, is then defined as

$$I_{\mathrm{p}} = |P + N|, \ (0 \leq I_{\mathrm{p}} \leq 2). \tag{2.17}$$

Smaller values of I_{p} indicate stronger push-pull effect.

2.3.10 Measuring contrast invariance of orientation tuning

The method in (Zhu and Rozell, 2013) was used to investigate contrast invariance of orientation tuning and the procedure is as follows. First, an exhaustive search was performed to find the preferred circular sinusoidal grating in the parameter space of the following ranges: radius of the grating was between 1 pixel and $2.5\min(\sigma_x, \sigma_y)$ (smaller than 8 pixels which is the maximum radius for a 16×16 image patch) with the stepsize of 1 pixel ; spatial frequency was between 0.05 and 0.3 cycles/pixel with the stepsize of 0.05 cycles/pixel; orientation was between 0 and 180 degrees with the stepsize of 5 degrees; phase was between 0 and 360 degrees with the stepsize of 30 degrees. Next, we measured the mean response to the drifting grating with orientations between 0 and 180 degrees with the stepsize of 5 degrees while varying the contrast of the stimuli from 0.2 to 1 in increments of 0.2, where contrast is defined as the amplitude of the sinusoidal grating. The orientation tuning curve for each contrast level was then fit to the Gaussian function and the half-height bandwidth of the Gaussian fit was calculated. The slope of the linear fit to half-height bandwidth vs. contrast for the cell was used to plot the population statistics of contrast invariance (Alitto and Usrey, 2004). Since we measured contrast invariance of orientation tuning, only oriented RFs inside the field were considered. After screening, 68 model simple cells that have oriented RFs located well within the 16×16 image patch were selected for the analysis.

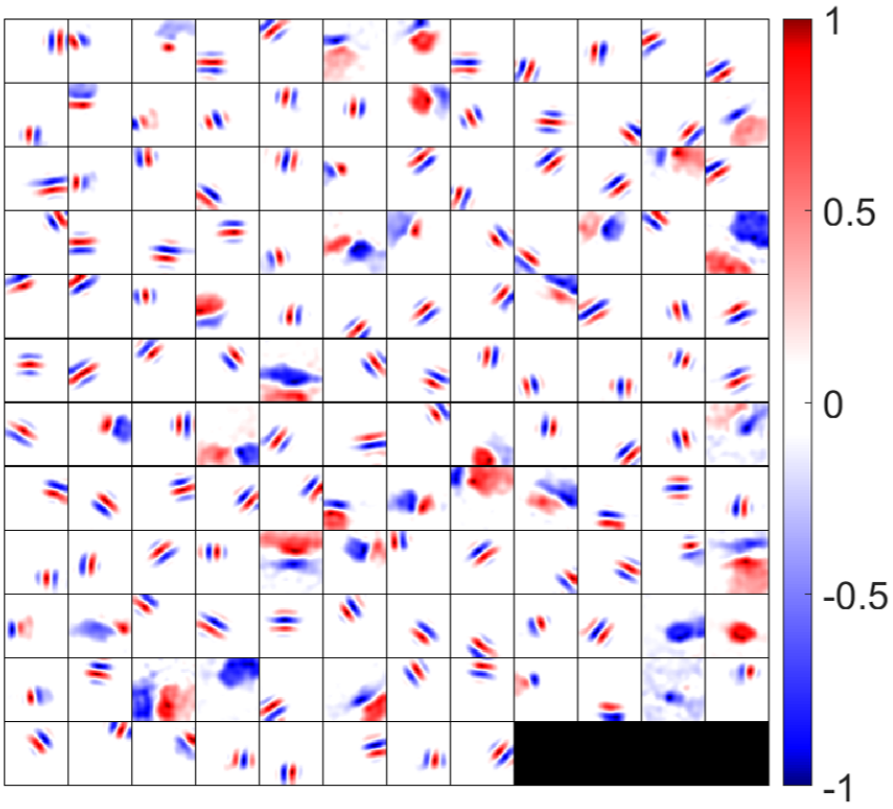

Figure 2.5: Synaptic fields (defined in Eq. 2.18) for 140 selected simple cells. Each block is a 16×16 image that represents the combined effects of ON and OFF LGN cells for a simple cell in spatial domain. 140 cells are located on a 12×12 grid. Values in each block are normalized to the range $[-1\ 1]$ when plotting this figure.

2.4 Results

After learning, synaptic weights between LGN and V1 display spatial structures similar to those observed in recordings of neurons in V1, such as oriented Gabor-like filters and non-oriented blobs. Since both excitatory and inhibitory connections from ON and OFF LGN cells contribute to the responses of simple cells, we use the *synaptic field* ($\mathbf{S}_\mathrm{f}$) defined as

$$\mathbf{S}_\mathrm{f} = (\mathbf{A}_\mathrm{ON}^{\mathrm{u},+} + \mathbf{A}_\mathrm{ON}^{\mathrm{u},-}) - (\mathbf{A}_\mathrm{OFF}^{\mathrm{u},+} + \mathbf{A}_\mathrm{OFF}^{\mathrm{u},-}) \tag{2.18}$$

to visualize the overall synaptic weights from ON and OFF LGN cells. The synaptic fields of 140 model simple cells that meet the two quality control measures (see the Materials and Methods section) are shown in Figure 2.5, where each block represents the overall effect of the feedforward connections from ON and OFF LGN cells to a simple cell. Note that although Figure 2.5 displays spatial patterns that are similar to experimental RFs, strictly they represent the synaptic weights from LGN cells to simple cells, which ignores the early visual processing before LGN. However, the RFs of the model are systematically investigated in the following sections.

In the remaining results, we show that the synaptic weights exhibit several properties that have been observed experimentally, including segregation of ON and OFF sub-regions, push-pull effect, phase-reversed feedback, diverse shapes of simple cell RFs, and contrast invariance of orientation tuning.

2.4.1 Segregated ON and OFF sub-regions

Hubel and Wiesel found that simple cells in cat striate cortex have spatially separated ON and OFF sub-regions (Hubel and Wiesel, 1962), which was also confirmed by other experimental studies (Jones and Palmer, 1987b; Hirsch et al., 1998; Martinez et al., 2005). However, it is impossible for a model that combines ON and OFF LGN input into a single linear input to explain this important phenomenon. Our model separates ON and OFF LGN input and shows that the learned feedforward excitatory connections from ON and OFF LGN cells to simple cells can account for the segregation of ON and OFF sub-regions of simple cells.

ON and OFF excitatory regions of some example simple cells are displayed in Figure 2.6A. In our model, there are 256 ON LGN and 256 OFF LGN cells located evenly on a 16×16 image, so each block in Figure 2.6A represents 256 excitatory weights from ON or OFF LGN cells to a simple cell. Figure 2.6A shows that these excitatory connections form spatial patterns such as bars and blobs. Furthermore, a careful examination of the patterns shows that excitatory connections from ON LGN cells are normally adjacent to patterns of excitatory connections from OFF LGN cells, but the two patterns do not overlap, as can be seen when contour plots for the ON and OFF excitatory regions are overlaid in Figure 2.6B.

We quantified the segregation of ON and OFF sub-regions using the overlap index (defined in the Materials and Methods section). The histogram of the overlap index for simple cells in an experimental study (Martinez et al., 2005) is re-plotted in Figure 2.6C. Consistent with the experimental data, 88 out of 92 model simple cells had an overlap index smaller than 0.1 (Figure 2.6D), which indicates that the ON and OFF sub-regions are well-separated in a large majority of the population. The synaptic fields of simple cells whose overlap indices are larger than 0.1 are shown in Figure 2.6E, revealing that most of them have low spatial frequencies.

2.4.2 Push-pull effect

Simple cells are also found to have push-pull responses; i.e., if one contrast polarity excites a cell, the opposite contrast polarity tends to inhibit it (Jones and Palmer, 1987b; Ferster, 1988; Hirsch et al., 1998; Martinez et al., 2005). Even though this effect has been observed in many experimental studies, to our knowledge there has not been a learning model

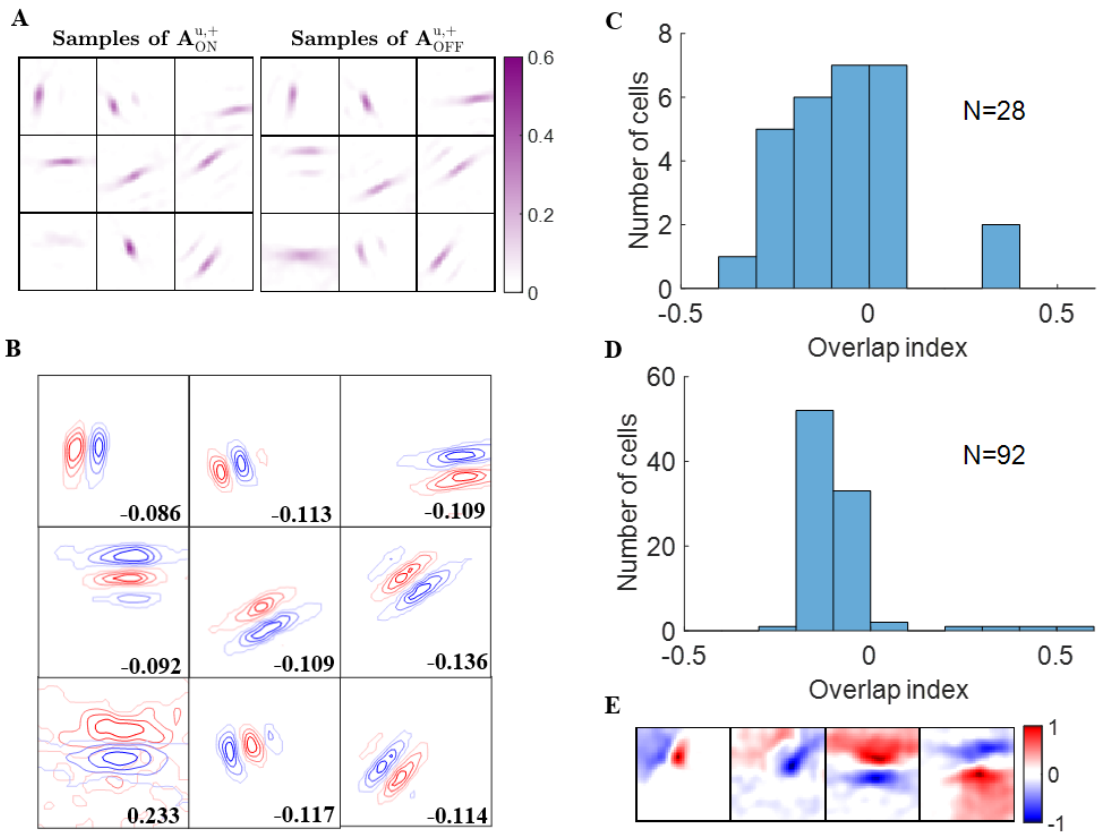

Figure 2.6: Segregation of ON and OFF sub-regions. (A) Some examples of $\mathbf{A}_{\text{ON}}^{\text{u},+}$ and $\mathbf{A}_{\text{OFF}}^{\text{u},+}$. Each block is a 16×16 image that represents 256 excitatory connections from ON or OFF LGN cells to a simple cell. The color magenta represents excitatory connections. (B) Red and blue contours represent excitatory connections from ON and OFF LGN cells, respectively. Connections that are smaller than 20% of the maximal connection were removed to only show the substantial weights. The number in each block indicates the overlap index. (C) Histogram of the overlap index for simple cells in cat V1. Re-plotted from Figure 3C in (Martinez et al., 2005). (D) Histogram of the overlap index for model simple cells. (E) Synaptic fields of the four simple cells with overlap index larger than 0.1.

proposed that can explain how this effect emerges. Again, a model that separates ON and OFF LGN input is necessary to investigate the emergence of the push-pull effect. In this section, we show that the push-pull effect for simple cells naturally emerges as a result of neural learning.

Some examples of ON excitatory and OFF inhibitory synaptic weights ($\mathbf{A}_{\text{ON}}^{\text{u},+}$ and $\mathbf{A}_{\text{OFF}}^{\text{u},-}$, respectively) are shown in Figure 2.7A. The patterns of $\mathbf{A}_{\text{ON}}^{\text{u},+}$ are similar to the ones of $\mathbf{A}_{\text{OFF}}^{\text{u},-}$ and they are located at similar locations, as can be seen from the highly overlapped contours in Figure 2.7B. However, the degree of overlap is different between the examples.

Analogous results to the above also hold for learned excitatory connections from OFF LGN cells, $\mathbf{A}_{\text{OFF}}^{\text{u},+}$, and inhibitory connections from ON LGN cells, $\mathbf{A}_{\text{ON}}^{\text{u},-}$ (data not shown).

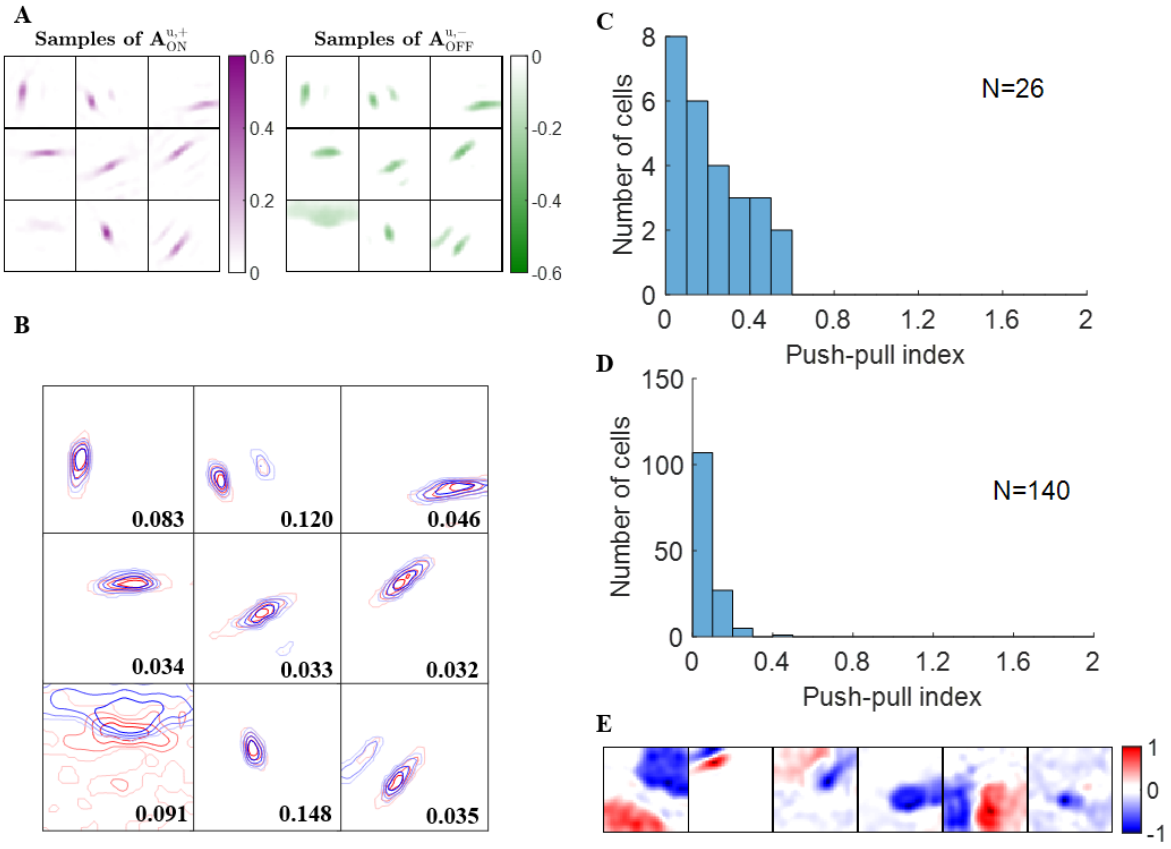

Figure 2.7: Push-pull effect. (A) Some examples of $\mathbf{A}_{\text{ON}}^{\text{u},+}$ and $\mathbf{A}_{\text{OFF}}^{\text{u},-}$. Each block on the left is a 16×16 image that represents 256 excitatory connections from ON LGN cells to a simple cell. Each block on the right represents inhibitory connections from OFF LGN cells to a simple cell. The color magenta represents excitatory connections; the color green represents inhibitory connections. (B) Red and blue contours represent excitatory connections from ON LGN cells ($\mathbf{A}_{\text{ON}}^{\text{u},+}$) and inhibitory connections from OFF LGN cells ($\mathbf{A}_{\text{OFF}}^{\text{u},-}$), respectively. Connections that are smaller than 20% of the maximal connection were removed to only show substantial weights. The number in each block indicates the push-pull index. (C) Histogram of the push-pull index for simple cells in cat V1. Re-plotted from Figure 4B in (Martinez et al., 2005). (D) Histogram of the push-pull index for model simple cells. (E) Synaptic fields of the six simple cells with push-pull index larger than 0.2.

We then quantified the push-pull effect using push-pull index (defined in the Materials and Methods section). Both the histograms of push-pull index for experimental data (Figure 2.7C) and model simple cells (Figure 2.7D) peaked near zero and showed an decreasing trend. Model simple cells showed even stronger push-pull index with more simple cells having push-pull index close to zero. The synaptic fields of simple cells with push-pull indices larger than 0.2 are shown in Figure 2.7E, showing that most of them have low spatial frequencies.

2.4.3 Phase-reversed feedback

The experimental study of Wang and colleagues suggests that the synaptic feedback from V1 to LGN is phase-reversed with respect to the feedforward connections (Wang et al., 2006). For example, the connection from a simple cell to an ON-center LGN cell will be excitatory if the ON-center is aligned in visual space to the OFF sub-field of simple cell (i.e., phase-reversed). Conversely, if the ON-center is aligned to the ON sub-field of the simple cell, the connection will be inhibitory. Our learning model with separate ON and OFF LGN cells enables us to investigate the feedback effect from simple cells to LGN cells. In this section, we show that phase-reversed feedback arises in the structures of learned connections.

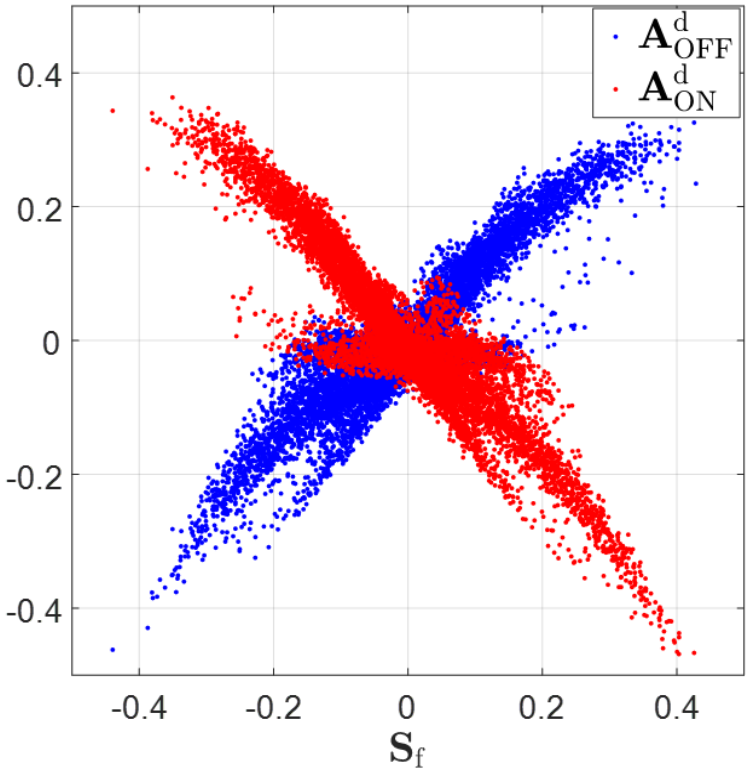

Figure 2.8: Synaptic fields, $\mathbf{S}_{\mathrm{f}}$ (defined in Eq. 2.18), vs. feedback to ON and OFF LGN cells, $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$ and $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}}$. $\mathbf{S}_{\mathrm{f}}$ is highly positively correlated with $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}}$ (correlation coefficient $r = 0.90$) and $\mathbf{S}_{\mathrm{f}}$ is highly anti-correlated with $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$ (correlation coefficient $r = -0.92$). When $\mathbf{S}_{\mathrm{f}}$ is greater than zero, $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}}$ tends to be greater than zero and $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$ tend to be smaller than zero. On the contrary, $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}}$ tends to be smaller than zero and $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$ tends to be greater than zero if $\mathbf{S}_{\mathrm{f}}$ is negative.

Feedback from simple cells to LGN cells occurs via both excitatory connections, $\mathbf{A}^{\mathrm{d},+}_{\mathrm{x}}$, and inhibitory connections, $\mathbf{A}^{\mathrm{d},-}_{\mathrm{x}}$, with the overall effect characterized by $\mathbf{A}^{\mathrm{d}}_{\mathrm{x}} = \mathbf{A}^{\mathrm{d},+}_{\mathrm{x}} + \mathbf{A}^{\mathrm{d},-}_{\mathrm{x}}$, where x = ON or OFF depending on the type of LGN cell. Therefore, the overall feedback to ON LGN cells, denoted as $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$, can be represented by $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}} = \mathbf{A}^{\mathrm{d},+}_{\mathrm{ON}} + \mathbf{A}^{\mathrm{d},-}_{\mathrm{ON}}$. Similarly, $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}} = \mathbf{A}^{\mathrm{d},+}_{\mathrm{OFF}} + \mathbf{A}^{\mathrm{d},-}_{\mathrm{OFF}}$ represents the overall feedback to OFF LGN cells.

The ON and OFF sub-fields of simple cells receptive fields are characterized by the positive and negative regions of the synaptic field defined in Eq. 2.18. The scatter plots in Figure 2.8 show that relationship expected from phase-reversed feedback. $\mathbf{S}_{\mathrm{f}}$ is highly positively correlated with $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}}$ (correlation coefficient $r = 0.90$), while $\mathbf{S}_{\mathrm{f}}$ is highly anti-correlated with $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$ (correlation coefficient $r = -0.92$). According to the figure, wherever $\mathbf{S}_{\mathrm{f}}$ is positive, indicating the ON sub-field, the feedback to ON LGN cells, $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$, is very likely to be negative and the feedback to OFF LGN cells, $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}}$, tends to be positive; however, wherever $\mathbf{S}_{\mathrm{f}}$ is negative, indicating the OFF-field, the converse is true: the feedback to ON LGN cells, $\mathbf{A}^{\mathrm{d}}_{\mathrm{ON}}$, is very likely to be positive and the feedback to OFF LGN cells, $\mathbf{A}^{\mathrm{d}}_{\mathrm{OFF}}$, tends to be negative. This corresponds to a phase-reversed feedback from V1 to LGN.

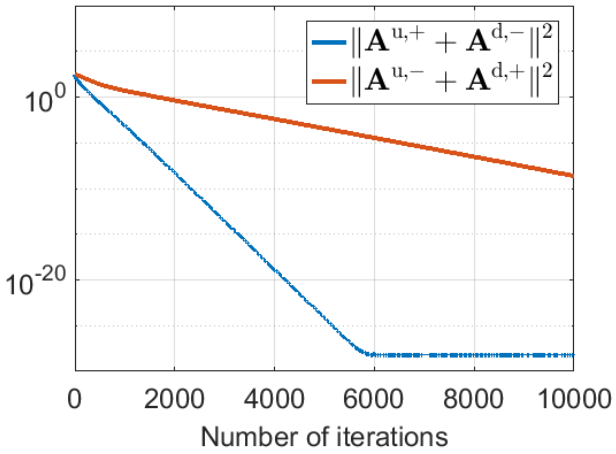

Figure 2.9: $\|\mathbf{A}^{\mathrm{u},+} + \mathbf{A}^{\mathrm{d},-}\|^2$ and $\|\mathbf{A}^{\mathrm{u},-} + \mathbf{A}^{\mathrm{d},+}\|^2$ during pre-development when white noise is used as the input. The difference between $\mathbf{A}^{\mathrm{u},+}$ and $-\mathbf{A}^{\mathrm{d},-}$ (blue line) decreases to zero very quickly during learning. Similarly, the difference between $\mathbf{A}^{\mathrm{u},-}$ and $-\mathbf{A}^{\mathrm{d},+}$ (red line) reduces to zero quickly, although somewhat slower than the blue line.

This phase-reversed feedback from V1 to LGN can be explained by the learning dynamics of LGN and simple cells described in Eqs. 2.8. The learning rule shows that $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{d},-}$ are updated with the same magnitude of synaptic change but opposite in sign (and are normalized with the same norm l_1). Similarly, $\mathbf{A}^{\mathrm{u},-}$ and $\mathbf{A}^{\mathrm{d},+}$ are updated with the same magnitude of synaptic change but opposite in sign (and are normalized with the same norm l_2). These anti-symmetries are a consequence of having Hebbian learning for the forward weights and anti-Hebbian learning for the feedback weights. In both cases the magnitude of weight change is proportion to the production of pre- and post-synaptic spike rates, but the sign of the change is opposite. The anti-symmetry arises because roles of pre- and post-synaptic rates are interchanged in forward versus feedback directions, in combination with the sign change. Simulation results show that $\mathbf{A}^{\mathrm{u},+}$ converges to $-\mathbf{A}^{\mathrm{d},-}$ and $\mathbf{A}^{\mathrm{u},-}$ converges to $-\mathbf{A}^{\mathrm{d},+}$ even during pre-development when white noise is used as

the input to the model, as illustrated in Figure 2.9.

2.4.4 The diversity of model receptive fields resembles that observed experimentally for simple cells

In this section, we show that the range of spatial structures of RFs of our model have a close resemblance to experimental data.

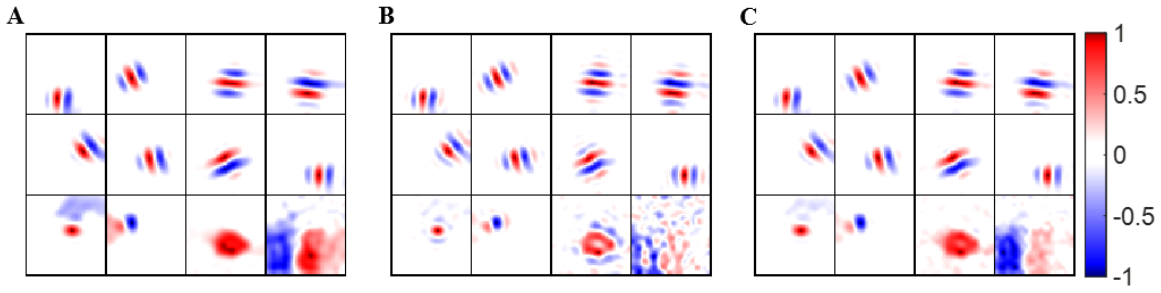

Figure 2.10: Receptive fields of example model cells. Values of each block are normalized to the range $[-1\ 1]$ when plotting the figure. (A) Synaptic fields of example model cells. (B) Pre-whitened RFs of example model cells. The pre-whitening filter described in Eq. 2.9 was used to filter white noise stimuli. (C) Low-pass RFs of example model cells. The low-pass filter described in Eq. 2.11 was used to filter white noise stimuli.

RFs were calculated from the model by simulating experiments in which Gaussian white noise is presented as a visual stimulus, and the spike triggered average is used to estimate RFs. As the presentation of white noise may cause adaptive effects in the early stages visual system relative to natural images, we considered two versions of the model, one with the standard pre-whitening filter (Eq. 2.9) modeling center-surround processing, and a second without pre-whitening in which the filter is replaced by a low-pass filter (Eq. 2.11) with the same upper cut-off frequency as pre-whitening filter. We use *pre-whitened RFs* and *low-pass RFs* to represent of simple cell RFs estimated using the pre-whitening filter and low-pass filter.

Some examples of pre-whitened RFs, low-pass RFs and synaptic fields are shown in Figure 2.10, which shows that pre-whitened RFs and low-pass RFs are similar to synaptic fields. However, pre-whitened RFs tend to have more and thinner stripes, which indicates a narrower tuning to a somewhat higher spatial frequency. For a simple cell tuned to very low spatial frequencies (bottom right blocks), the RF recovered with pre-whitening was a poor match to the original synaptic field, but for RF recovered with low-pass filtering it was fair.

Early studies show that RFs of simple cells can be well-described by 2D Gabor functions described in Eq. 2.12 (Jones and Palmer, 1987a; Ringach, 2002). For our model, most RFs

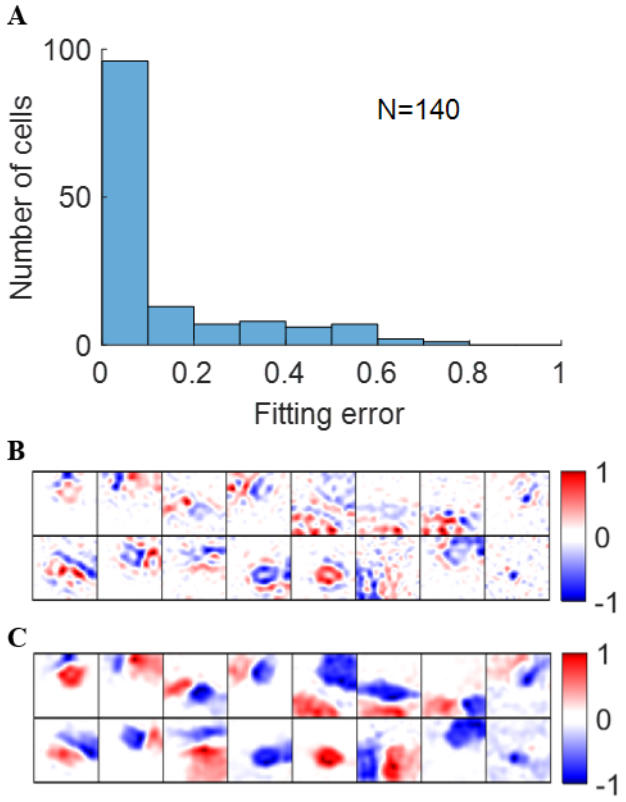

Figure 2.11: (A) Histogram of Gabor fitting errors for pre-whitened RFs. (B) Pre-whitened RFs that has fitting error larger than 40%. (C) Synaptic fields of the corresponding cells in (B).

could be well-fitted by Gabor functions with suitable choices of parameters with small fitting errors, as shown in Figure 2.11A. Note that although the fitting error of blob-like RFs might be low, the parameter choices are not necessarily reasonable, in that they are poorly constrained and the process of Gabor fitting imposes an a priori hypothesis that the RF is a 2D-Gabor function even though it is clearly not Gabor-like. The pre-whitened RFs with fitting errors larger than 40% (Figure 2.11B) are cells whose synaptic fields have low spatial frequencies (Figure 2.11C), because pre-whitened RFs of these cells matched poorly to the original synaptic fields (Figure 2.10B). Low-pass RFs of all 140 selected model cells have fitting errors smaller than 40% with 132 of them having fitting errors smaller than 20% (data not shown).

Using fitted parameters of Gabor functions, Ringach constructed a scatter plot of $n_x = \sigma_x f_s$ vs. $n_y = \sigma_y f_s$ to analyze the spatial structures of RFs in V1 over the population (Ringach, 2002). Such plots have subsequently been used by many researchers to investigate the distributions of model simple cell RFs (Rehn and Sommer, 2007; Zylberberg et al., 2011; Wiltschut and Hamker, 2009). n_x and n_y are the width and length of the Gabor function measured in the number of cycles of the spatial frequency (i.e., across and along

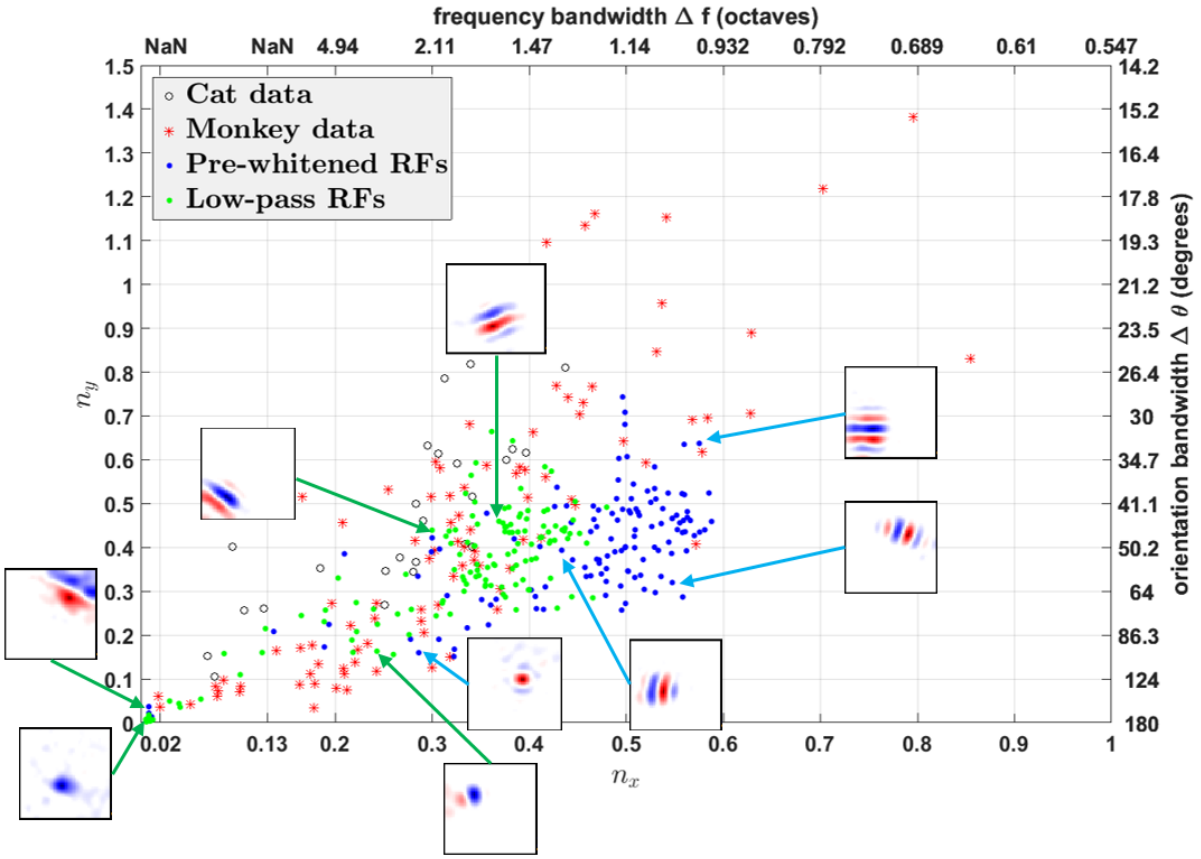

Figure 2.12: n_x vs. n_y. Comparison of RFs of the model with experimentally recorded data for cat simple cells and monkey simple cells. Open circles: 25 cat simple cells from Table 1 in (Jones and Palmer, 1987a) re-plotted in the (n_x, n_y) plane; red stars: 93 monkey simple cells in (Ringach, 2002); blue dots: pre-whitened RFs using the pre-whitening filter described in Eq. 2.9; green dots: low-pass RFs using the low-pass filter described in Eq. 2.11. The axes on the top and right represent frequency and orientation bandwidths of fitted Gabor functions computed using Eq. 2.19. Some examples of RFs are displayed in the inset sub-plots. Data points of estimated RFs with fitting errors $> 40\%$ were excluded, which gave 124 data points for pre-whitened RFs and 140 data points for low-pass RFs.

the stripes). Ringach noted that blob-like RFs are mapped to points near the origin, while RFs with elongated sub-regions are mapped to points away from the origin (Ringach, 2002). In addition, n_x and n_y are directly related with the half-magnitude spatial frequency bandwidth Δf and orientation bandwidth $\Delta\theta$ of the fitted Gabor function,

$$\begin{aligned} \Delta f &:= h(n_x) = \log_2 \left(\frac{1 + \frac{\sqrt{2\ln 2}}{2\pi n_x}}{1 - \frac{\sqrt{2\ln 2}}{2\pi n_x}} \right) \text{ in octaves} \\ \Delta\theta &:= g(n_y) = 2 \arctan \left(\frac{\sqrt{2\ln 2}}{2\pi n_y} \right) \text{ in degrees.} \end{aligned} \tag{2.19}$$

Both $h(n_x)$ and $g(n_y)$ are monotonically decreasing functions; i.e., the larger n_x and n_y, the

smaller Δf and $\Delta\theta$. Note that $h(n_x)$ is not well defined when $n_x < \sqrt{2\ln 2}/2\pi$ (≈ 0.13), i.e., when the lower half-magnitude frequency do not exist. This corresponds to the region in which Gabor fitting gives ambiguous fits for parameters like spatial frequency and orientation, because oriented RFs with low spatial frequency might lie in this region as well.

We plot n_x vs. n_y and Δf vs. $\Delta\theta$ for RFs obtained from both the model and experimental studies in Figure 2.12. However, the different pre-processing filters for white noise stimuli have a dramatic influence on the distributions of n_x vs. n_y, shifting the distribution for low-pass RFs to the left of pre-whitened RFs, in closer agreement to the experimental data. As mentioned earlier, pre-whitened RFs tend to have more stripes relative to the low-pass RFs, so they are mapped to points away from the origin compared to low-pass RFs. In addition, the distribution of low-pass RFs is continuous from the origin, while there is a gap between points near the origin and points away from the origin for pre-whitened RFs. The inset sub-plots of Figure 2.12 show that data points near the origin might be orientated RFs with low spatial frequencies and blob-like RFs might not be necessarily mapped to points near the origin.

In general, oriented RFs are well described by Gabor functions and low-pass RFs better resemble the distribution of experimental data compared with pre-whitened RFs.

2.4.5 Contrast Invariance of Orientation Tuning

Another important property of simple cells is contrast invariance of orientation tuning; i.e., the width of the orientation tuning curve is maintained when the contrast of the stimulus changes, as demonstrated in Figure 2.13A. The orientation tuning curves with various stimulus contrasts for a model simple cell are shown in Figure 2.13B, where the bandwidths of each curve remain the same while the responses become larger when the stimulus contrast increases. For a study of contrast invariance of V1 population in ferret, the histogram of the slope of the linear fit of half-width bandwidth vs. contrast (Figure 2.13C) showed that most cells were contrast invariant with the slope close to zero (Alitto and Usrey, 2004). Figure 2.13D shows that most model cells have the slope around zero, which is consistent with experimental data.

Contrast invariance of orientation tuning for simple cells has also been explained by models in two previous studies (Banitt et al., 2007; Kremkow et al., 2016). Banitt et al. (2007) applied a neural adaptation in which thalamocortical synapse can depress as a function of frequency and Kremkow et al. (2016) used a combination of synaptic depression with push-pull inhibition that acts as a feedforward-like inhibition. The model designed in this chapter introduces competition into the network by the feedforward-feedback loop, adjusting and compressing the neuronal responses, which is similar to the function of synaptic depression used in both these studies. In addition, the learned model displays

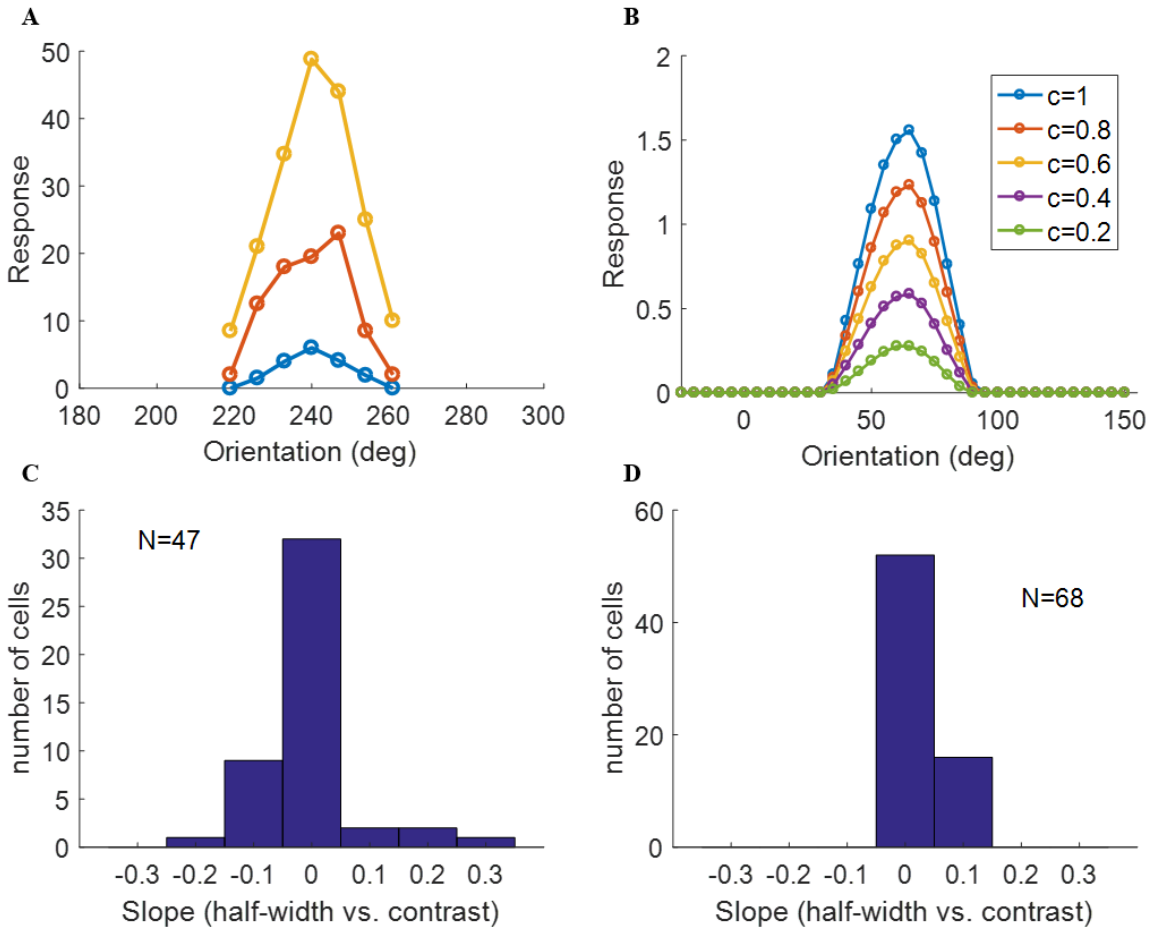

Figure 2.13: Contrast invariance of orientation tuning. (A) Contrast invariant orientation tuning curves of a simple cell in cat V1. Re-plotted from Figure 3A in (Skottun et al., 1987). Different colors represent different contrasts. (B) Contrast invariant orientation tuning curves of a cell in our model. $c = 1$ and $c = 0.2$ correspond to the high and low contrast, respectively. (C) Histogram of the slope of half-height bandwidth vs. contrast for V1 population in ferret. Re-plotted from Figure 3B in (Alitto and Usrey, 2004). (D) Histogram of the slope of half-height bandwidth vs. contrast for model simple cells.

push-pull effect, which is similar to the feedfoward-like inhibition in Kremkow et al. (2016).

2.5 Discussion

2.5.1 Relationship with sparse coding

Sparse coding has been successful in modelling simple cell receptive fields (RFs) and has been used by many researchers over the past years. Our model is based on an algorithm that efficiently implements sparse coding (Rozell et al., 2008), and is therefore closely related to the original concept of sparse coding (Olshausen and Field, 1996).

If we define $\mathbf{A}$ as a $2N \times M$ matrix that represents the overall effect caused by

excitatory and inhibitory connections from $2N$ LGN cells to M simple cells, we have $\mathbf{A} = \mathbf{A}^{\mathrm{u},+} + \mathbf{A}^{\mathrm{u},-}$. The dynamics of simple cells described in Eq. 2.7 can be rewritten as

$$\tau_{\mathrm{C}}\dot{\mathbf{v}}^{\mathrm{C}} = -\mathbf{v}^{\mathrm{C}} + \mathbf{A}^{T}(\mathbf{s}^{\mathrm{L}} - s_{\mathrm{b}}) + \mathbf{s}^{\mathrm{C}}. \tag{2.20}$$

As illustrated in Figure 2.9, $\mathbf{A}^{\mathrm{u},+} \rightarrow -\mathbf{A}^{\mathrm{d},-}$ and $\mathbf{A}^{\mathrm{u},-} \rightarrow -\mathbf{A}^{\mathrm{d},+}$ during learning. Therefore, we have $\mathbf{A}^{\mathrm{d},-} + \mathbf{A}^{\mathrm{d},+} = -\mathbf{A}^{\mathrm{u},+} - \mathbf{A}^{\mathrm{u},-} = -\mathbf{A}$. The dynamics of LGN cells described in Eq. 2.5 can be rewritten as

$$\tau_{\mathrm{L}}\dot{\mathbf{v}}^{\mathrm{L}} = -\mathbf{v}^{\mathrm{L}} + \mathbf{x} - \mathbf{A}\mathbf{s}^{\mathrm{C}} + s_{\mathrm{b}}. \tag{2.21}$$

If the columns of $\mathbf{A}$ are seen as the basis vectors of a generative model, $\mathbf{A}\mathbf{s}^{\mathrm{C}}$ can be seen as the linear reconstruction of the input using learned basis vectors and thus $\mathbf{x} - \mathbf{A}\mathbf{s}^{\mathrm{C}}$ represents the residual error, which is similar to $\mathbf{r}$ of the sparse coding formulation given in Eq. 2.2. Therefore, the residual error used to update the basis vectors of the original sparse coding model is represented by the responses of LGN cells in our model.

To incorporate Dale's law, non-negative connections, $\mathbf{A}^{\mathrm{u},+}$, and non-positive connections, $\mathbf{A}^{\mathrm{u},-}$, are employed in our model to represent the positive and negative elements of $\mathbf{A}$. $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{u},-}$ are not co-active in general, which suggests that $\mathbf{A}^{\mathrm{u},+} \approx [\mathbf{A}]_{+}$ and $\mathbf{A}^{\mathrm{u},-} \approx [\mathbf{A}]_{-}$, where $[\,\cdot\,]_{+}$ preserves the positive elements and sets negative elements to zero and $[\,\cdot\,]_{-}$ preserves the negative elements and sets positive elements to zero.

In other words, our model is essentially a variant of sparse coding that employs separate connections to learn the positive and negative part of the overall connections.

2.5.2 Relationship with predictive coding

Our model is a hierarchical model with feedforward and feedback connections based on a locally competitive algorithm (Rozell et al., 2008). The structure of our model is essentially very similar to that of predictive coding models. To be more specific, the feedback from the second-layer neurons reconstructs the input. The residual error is computed at the first layer and then propagated to the second layer via feedforward connections.

Although our model presented here and the predictive coding model of Jehee and Ballard (Jehee and Ballard, 2009) can explain phase-reversed feedback, the models differ in several respects. First, sparse coding in our model is simply realized by the threshold of the rectifying function of firing rates for simple cells and this simple mechanism leads to simple neural circuits. Second, compared to the mechanism for determining simple cell responses one by one in their model, our model computes the responses in parallel. Third, our model generates diverse types of RFs that correspond well to experimental data. Finally, the phase-reversed effect is simply accounted for by the special pattern of learned connections, which also explains the segregation of ON/OFF sub-regions and

push-pull effect for simple cells.

2.5.3 The Function of spontaneous activity

In the model proposed here, the dynamics of LGN cells described in Eq. 2.5 has the background firing rate, s_{b}, as part of the input to LGN cells. This spontaneous firing rate introduces a shift of the operating point for LGN cells. Given the responses of simple cells, $\mathbf{s}^{\mathrm{C}}$, $\mathbf{x} - \mathbf{A}\mathbf{s}^{\mathrm{C}}$ in Eq. 2.21 represents the reconstruction residual error between the input and reconstruction. The residual error gives the difference between the real input and the representation produced by the model and it can be either positive or negative. To code for the signed quantities (residual error), Ballard and Jehee carried out a case-by-case study, leading to very complicated neural circuits (Ballard and Jehee, 2012). However, our model has a straightforward method for the implementation of solving signed quantities. The background firing rate, s_{b}, in Eq. 2.5 increases the residual errors by s_{b}. Therefore, the membrane potential of LGN cell, $\mathbf{v}^{\mathrm{L}}$, represents the residual error shifted up by s_{b}. The threshold function in Eq. 2.5 gives the firing rate of the LGN cell and it preserves the residual error in the interval of $[-s_{\mathrm{b}},\ \infty]$, which preserves the information of whether the model under-estimates or over-estimates the input stimuli and forces the connections to evolve through learning in the correct direction. In Eq. 2.7, which describes simple cell dynamics, the effect of the spontaneous firing rate, s_{b}, is removed by $\mathbf{v}^{\mathrm{C}}_{\mathrm{leak}}$, a homeostatic mechanism employed by simple cells to maintain resting membrane potentials when there is no external input. The local learning rule described by Eq. 2.8 also eliminates the effect of the spontaneous firing rate by subtracting it. The use of spontaneous firing rate makes the model much simpler and offers a new approach for solving the problem of signed quantities (residual errors). Experimental evidence shows that thalamocortical neurons can fire with bursts of action potentials without any synaptic input (Kandel et al., 2013), which suggests that the spontaneous firing activities might be used to encode the difference between input and feedback information.

2.5.4 Pre-processing of the early visual system

Atick and Redlich suggest that the retinal goal is to whiten the visual input up to a transition frequency such that input noise can also be suppressed (Atick and Redlich, 1992). The pre-whitening filter (Eq. 2.9) approximately whitens the natural scenes up to the cut-off frequency.

However, for pre-processing white noise stimuli, two hypotheses are considered here. First, the filtering process of the early visual system can be described by the pre-whitening filter (Eq. 2.9) whether or not the visual stimuli are natural scenes. Second, the early visual system is adaptive such that the visual stimuli are whitened up to a cut-off frequency.

In this case, a low-pass filter (Eq. 2.9) should be used, because white noise stimuli are already whitened across all frequencies. Our results suggest that estimated RFs using low-passed white noise match the experimental data much better than estimated RFs using pre-whitened white noise. Further investigation of how visual stimuli are processed before they are fed to the visual cortex is needed to better understand the properties of simple cells.

2.5.5 The role of l_1 and l_2

Each column of $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{d},-}$ is normalized to norm l_1 and each column of $\mathbf{A}^{\mathrm{u},-}$ and $\mathbf{A}^{\mathrm{d},+}$ is normalized to norm l_2. In other words, l_1 represents the overall strength of feedforward excitatory connections and feedback inhibitory connections while l_2 represents the overall strength of feedforward inhibitory connections and feedback excitatory connections. The results shown in this chapter are based on $l_1 = 1$ and $l_2 = 1$; i.e., the strength of feedforward excitatory connections is equivalent to feedforward inhibitory connections, which leads to a strong push-pull effect in Figure 2.7D. If l_2 is smaller than l_1, the push-pull effect will be weaker and the distribution of the push-pull index will shift to the right. In addition, reducing l_2 results in more blob-like receptive fields (data not shown).

2.5.6 Neural circuits

Biologically realistic neural models can provide deeper insights into how real neural circuits function. The model proposed here contains a number of features that correspond to those in its biological counterpart, namely in terms of ON and OFF channels for LGN cells, positive neuronal responses, local computation, local learning rule, existence of feedback, and obedience to Dale's law.

In addition, our model incorporates inhibitory effects between LGN cells and cortical simple cells. As pointed out in the Materials and Methods section, for simplicity, inhibitory effects are implemented by direct inhibitory connections between two layers. However, in reality, long-range inhibitory effects should be implemented via interneurons that have inhibitory synapses. In this section, we will discuss several neural circuits of implementing inhibitory connections of our model.

Possible neural circuits that may be used to implement long-range inhibition are displayed in Figure 2.14. Assume that the overall inhibitory effects from LGN cells (with activity $\mathbf{s}^{\mathrm{L}}$) to cortical simple cells (with activity $\mathbf{s}^{\mathrm{C}}$) can be represented by inhibitory connections, $\mathbf{A}^{-}$, between populations. We also assume that the learning rule of $\mathbf{A}^{-}$ is local, i.e., that only depends on the responses of two populations ($\mathbf{s}^{\mathrm{L}}$ and $\mathbf{s}^{\mathrm{C}}$). Long-range inhibition in our model is implemented via direct inhibitory connections, which is not biologically realistic (Figure 2.14A).

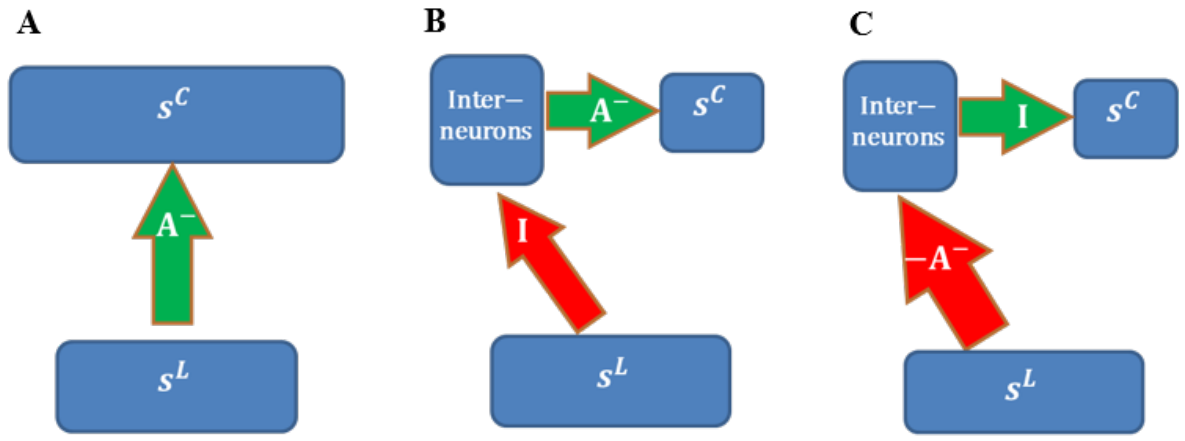

Figure 2.14: Possible neural circuits for implementing long-range inhibition. Red and green arrows represent excitatory and inhibitory connections. (A) Direct long-range inhibition. (B) Circuit I. (C) Circuit II.

The circuit in Figure 2.14B implements inhibitory connections, $\mathbf{A}^-$ (with non-positive weights), via a population of interneurons that have inhibitory connections, $\mathbf{A}^-$, with cortical simple cells. LGN cells are connected to interneurons via long-range identical excitatory connections, $\mathbf{I}$; i.e., the interneurons copy the responses of LGN cells. For this structure, long-range excitatory connections, $\mathbf{I}$, are fixed while $\mathbf{A}^-$ are learned using the same learning rule in Figure 2.14A. In this case, the learning rule of $\mathbf{A}^-$ is still local because the responses of interneurons are just $\mathbf{s}^L$ and the model is still biologically plausible in terms of the local learning rule. Furthermore, the RFs of interneurons in the same layer as cortical simple cells should be LGN-like. Though V1 cortical cells with blob-like RFs were found in different species (Jones and Palmer, 1987a; Ringach, 2002; Chapman and Stryker, 1993; Hawken et al., 1988; Kretz et al., 1986; Muly and Fitzpatrick, 1992), we are not sure whether this neural circuit is the most likely candidate because the fixed identical connection between LGN cells and the interneurons seems artificial unless they can be learned.

Figure 2.14C shows another possible neural circuit for implementing $\mathbf{A}^-$. LGN Cells are connected to interneurons via long-range excitatory connections, $-\mathbf{A}^-$. There is a one-to-one mapping between interneurons and cortical simple cells. In this case, the overall effect from LGN cells to simple cells is equivalent to $\mathbf{A}^-$. In addition, the RFs of inhibitory interneurons should resemble simple cells and show orientation tuning since the learned $\mathbf{A}^-$ has spatial structures such as oriented bars, which is consistent with the smooth simple cells found in cat V1 of the experimental study (Hirsch et al., 2003). The positive connections $-\mathbf{A}^-$ can be learned by Hebbian learning and the identical connections between interneurons and cortical simple cells can be learned by anti-Hebbian learning. Therefore, this neural circuit is more feasible than than the circuit in Figure 2.14B.

2.5.7 Discrepancies between model and experimental data

Our model can capture the most significant features of experimental phenomena such as the segregation of ON and OFF sub-regions, push-pull effect and contrast invariance of orientation tuning. However, there are also discrepancies between the distributions of model and experimental data. In general, the histograms of experimental data (Figure 2.6C, Figure 2.7C and Figure 2.13C) are wider than model data (Figure 2.6D, Figure 2.7D and Figure 2.13D), which shows that experimental data is more diverse. One possible explanation is that model cells in this chapter are only a subset of the rich repository of real cortical cells. Furthermore, choices of free parameters in the model might also lead to different results.

2.6 Conclusion

In this chapter, we presented a biologically plausible model of LGN-V1 pathways to account for many experimental phenomena of V1. We found that the segregation of ON/OFF sub-regions of simple cells, push-pull effect, and phase-reversed cortico-thalamic feedback can all be explained by the structure of learning connections when the model incorporates ON and OFF LGN cells and is trained using natural images. Furthermore, the model can produce diverse shapes of receptive fields and contrast invariance of orientation tuning of simple cells, consistent with experimental observations.

Chapter 3

Can the principle of efficient coding learn complex cells?

3.1 Introduction

About 60 years ago, Hubel and Wiesel identified two distinct types of cells in the primary visual cortex (V1) of cat: simple cells and complex cells (Hubel and Wiesel, 1959, 1962). They categorized simple cells as cells that have a receptive field (RF) with distinct light (ON) and dark (OFF) regions, exhibit summation within ON and OFF regions and show antagonism between ON and OFF regions. Simple cells respond primarily to oriented edges with a strong preference for a particular orientation.

In contrast, complex cells exhibit significant non-linear spatial integration and do not show the above characteristics of simple cells. One important feature of complex cells is spatial phase invariance; i.e., they evoke strong responses to oriented bars with the preferred orientation for a wide range of spatial phases. Spatial phase invariance is similar to shift invariance or position invariance, which means that the response is generally not sensitive to the relative position of the stimulus within the RF of a complex cell. Movshon, Thompson and Tolhurst (1978a; 1978b) found that simple and complex cells have different degrees of response modulation when presented with drifting gratings. Subsequently, the degree of response modulation was defined by the ratio F_1/F_0 (De Valois et al., 1982), where F_1 is the component of the response to the drifting grating at the temporal drifting frequency, and F_0 is the DC component of the response, i.e., the mean response over time to the drifting grating with spontaneous activity subtracted. Cells are identified as complex if $F_1/F_0 < 1$ and simple if $F_1/F_0 > 1$. Skottun and colleagues found a bimodal distribution of the F_1/F_0 ratio in the population of cells in both cat and monkey V1 (Skottun et al., 1991).

3.1.1 Phenomenological models of complex cells

Simple cell responses can be described phenomenologically using a three-stage process (Movshon et al., 1978b; Carandini et al., 2005; Carandini, 2006): first, the input image is linearly filtered by a single filter whose weights represent the cell's receptive field; second, the feature contrast, the output of the first stage, is passed through a static nonlinearity (usually one-sided, e.g., half-wave rectification) to obtain the firing rate; third, spike trains are generated via a Poisson process. This is referred to a the linear-nonlinear-Poisson model.

A similar model with the three-stage process can be used to describe responses of complex cells, except that complex cells have multiple linear filters and the static nonlinearity might be double-sided (e.g., squaring nonlinearity) (Movshon et al., 1978a; Carandini et al., 2005; Carandini, 2006). One classical phenomenological model for complex cells is the energy model, where the complex cell responses are the summation of the squared outputs of two linear spatial-phase-shifted filters, as shown in Figure 3.1 (Adelson and Bergen, 1985; Carandini et al., 2005); the squaring nonlinearity brings polarity invariance (non-selective to the polarity of the stimulus) and filters with different spatial phases generate complex cells tuned to a wider range of spatial phases. Although the energy model can capture the spatial phase invariance of complex cells, a recent experimental study by Almasi and colleagues showed that complex cells in cat visual cortex have great diversity and only a minority can be characterized by the energy model (Cloherty and Ibbotson, 2014; Meffin et al., 2015; Almasi, 2017; Yunzab et al., 2019).

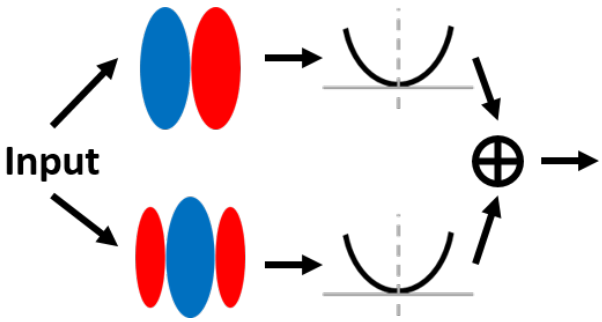

Figure 3.1: The energy model of a complex cell. Red and blue represent ON and OFF sub-regions of the receptive field of a simple cell, respectively. The linear response of the input convolved with the filter is passed to a two-side nonlinear function (a power function is this case). The outputs of two nonlinear functions are then summed to generate the response for the complex cell.

3.1.2 Computational models of complex cells

Hierarchical, parallel and recurrent structures

Computational models of the underlying mechanism of the formation of complex cells can be divided into three categories: hierarchical, parallel and recurrent (see Martinez and Alonso, 2003 for a review). The notion of the hierarchical model was proposed by Hubel and Wiesel (1962), where a complex cell pools the activities of simple cells of the same orientation but with different spatial phase preferences so that it is orientation selective but invariant to different spatial phases. The pooled simple cells form the *subspace* of the complex cell and each pooled simple cell is a *subunit* in the subspace. This idea was later supported by an experimental study conducted by Movshon et al. (1978a). From the perspective of a hierarchical structure, the energy model can be understood as a complex cell having convergent input from four simple cells with different spatial phase preferences; i.e., the response of the complex cell is the weighted sum of four simple cells in a *subspace* (Figure 3.2).

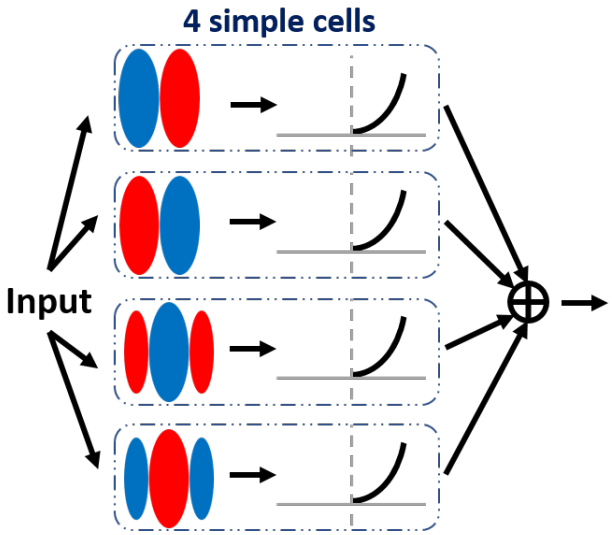

Figure 3.2: The equivalent hierarchical structure of the energy model of a complex cell. The response of the complex cell sums over the responses of simple cells. The colors red and blue represent ON and OFF sub-regions of the receptive fields of simple cells, respectively.

The concept of the hierarchical structure has been challenged by parallel and recurrent models. In the parallel model of Hoffmann and Stone (1971), it is proposed that both simple and complex cells are generated by separate thalamocortical pathways in parallel. This is supported by the discovery that some complex cells receive direct input from the thalamus (Hoffmann and Stone, 1971). However, some studies show that most complex cells do not receive direct input from the thalamus (Callaway, 2001; Martinez and Alonso, 2001; Martin, 2002). The idea of the recurrent model is strongly supported

by the experimental evidence that cortical cells mainly receive excitatory input from other cortical areas instead of the thalamus (Peters and Payne, 1993). Therefore, the response of cortical cells should be primarily determined by recurrent cortical inputs (Martin, 2002). Nevertheless, thalamocortical connections have many features to make them strong although they only account for a small fraction of excitatory synapses made by cortical cells (Martinez and Alonso, 2003).

Existing learning models of complex cells with a hierarchical structure

In this chapter, we focus on the hierarchical structure between simple and complex cells, namely that complex cells receive feedforward input from simple cells, as supported by experimental results (Van Kleef et al., 2010).

Given simple cells with RFs of different orientation and spatial phase preferences, there are two distinct mechanisms of pooling simple cells to construct complex cells: *indiscriminate pooling* and *selective pooling*.

Indiscriminate pooling means simply pooling all local simple cells from a narrow region of the map. If simple cells with similar orientation but different spatial phases are located in the local region, complex cells will be spatial-phase-invariant by indiscriminate pooling of these simple cells. Orientation maps are prevalent among monkey, cat and ferret V1 (Bartfeld and Grinvald, 1992; Ohki et al., 2006; Chenchal Rao et al., 1997), and some models were designed to describe complex cells based on the orientation topography of simple cells (Hyvärinen and Hoyer, 2001; Ma and Zhang, 2007; Antolik and Bednar, 2011). However, there are also rodent species, such as mouse and rat, that do not have orientation maps but still have complex cells in V1 (Bonin et al., 2011; Ohki et al., 2005).

Selective pooling means pooling simple cells according to certain criteria. For example, only simple cells with receptive fields (RFs) of the same orientation but different spatial phase preferences can be pooled. Selective pooling seems to be a common principle for constructing complex cells by pooling simple cells. Nevertheless, the questions remain unclear of how to pool simple cells, which simple cells should be pooled and how strong the pooling weights should be, and most existing models overlook many details of biological reality.

Hyvärinen and Hoyer proposed a model called *independent subspace analysis* (ISA) using the principle of *independent component analysis* (ICA) (Comon, 1994) to find independent feature subspaces of natural images (Hyvärinen and Hoyer, 2000). After training the model using static natural images, the subunits in each subspace are linear filters with similar orientation but different spatial phases, and the response of each subspace is invariant to spatial phase and position changes. Their model assumes a squaring nonlinearity and the weights for each subunit are pre-fixed; the learned subspace resembles the energy model if the number of subunits in each subspace is set to two

(Hyvärinen and Hoyer, 2000). If the subspace and subunits within ISA represent a complex cell and simple cells pooled by this complex cell, there are two significant problems if the model is implemented in a biological neural system: 1) the squaring nonlinearity in ISA is unrealistic for a neuron to have such a firing mechanism; 2) the fixed weights connecting subunits of ISA are biologically unrealistic unless the value of fixed weights can be learned via a biologically plausible learning rule.

Berkes and Wiskott used *slow feature analysis* (SFA) to investigate temporal slowness as a learning principle for RFs, an algorithm that determines functions that can extract slowly varying component from the input (Berkes and Wiskott, 2005). In order to incorporate temporal information from static images, the model was trained using translated sequences of static images as the input. After training, the functions determined by the algorithm could explain many complex cell properties (Berkes and Wiskott, 2005). However, SFA does not specify any direct biological realization. First, the focus of SFA is not biological plausibility. Second, SFA algorithm only solves an optimization problem and implies no Hebbian plasticity between units of a network. Furthermore, the functions have the square computation and assume the model units have a squaring nonlinearity to implement it, which is not realistic for the firing mechanism of biological neurons.

Einhäuser et al. (2002) designed a model with a learning rule inspired by the trace rule (Földiák, 1991), a principle that the response is determined by the trace (history) of activities, to learn the weights between simple and complex cells. After the model was trained on natural videos, model complex cells could pool simple cells with similar orientations but different positions, which gave complex cells spatial phase invariance (Einhäuser et al., 2002). However, the fact that only a winner can learn in each iteration is artificial because it requires the network to have global information to detect the winning neuron and incorporate a mechanism that shuts down the learning process of other synapses not connected with the winner. Furthermore, the ratio of simple to complex cells, $60:4$, is much larger than experimental evidence that complex cells are at least as prevalent as simple cells in V1 (Hubel and Wiesel, 1968). In addition, the learned RFs of simple cells do not match well with experimental results.

Hosoya and Hyvärinen (2016) applied strong *principle component analysis* (PCA) dimension reduction to pooling simple cell RFs trained on natural images using standard ICA. Their approach can pool reasonable subunits for complex cells, but the weights connecting simple and complex cells are not learned and thus it is not clear how this model can be implemented in a biological system.

3.1.3 Aim of this chapter

The principle of efficient coding finds an efficient representation of the sensory input by minimizing the average activity of model units (details in Section 2.3) and has been

successful in generating simple cell RFs (Olshausen and Field, 1996, 1997; Rehn and Sommer, 2007; Wiltschut and Hamker, 2009; Zylberberg et al., 2011). When more biological constraints are incorporated to build a more biologically plausible model of the LGN-V1 pathway based on efficient coding, the model can account for other experimental phenomena, such as the segregation of ON and OFF sub-regions, the push-pull effect and phase-reversed feedback (Chapter 2, Lian et al. (2019)). However, whether complex cells can be learned in a biologically plausible model based on efficient coding is still not clear. Here, we investigate whether efficient coding can be used to learn connections between simple and complex cells.

3.2 Methods

This model is based on the principle of efficient coding that finds a linear representation of the input using a small number of model units (see Chapter 2 for details).

3.2.1 Structure of the model

A four-layer model was built to describe the activities of lateral geniculate nucleus (LGN) cells (first layer), V1 simple cells (second layer), intermediate cells (third layer), and V1 complex cells (fourth layer). The intermediate cells take simple cell responses and provide input for complex cells, which separates the computations of simple and complex cells, and helps us investigate efficient coding for complex cells much easier. However, the model implies no structure of biological neural circuits. Model symbols and parameters used throughout the chapter are given in Table 3.1.

The bottom two layers implement the two-layer model (graphical representation shown in Figure 3.3A) that is biologically plausible and can account for many experimental phenomena (see Chapter 2, Lian et al. (2019)). This two-layer model is a variant of sparse coding and incorporates many biological constraints such as Dale's Law, non-negative firing rates and separate connections. The dynamics of LGN cells and simple cells are given by

$$\begin{aligned} \tau_{\mathrm{L}}\dot{\mathbf{v}}^{\mathrm{L}} &= -\mathbf{v}^{\mathrm{L}} + \mathbf{x}^{\mathrm{L}} + (\mathbf{A}^{\mathrm{d},+} + \mathbf{A}^{\mathrm{d},-})\mathbf{r}^{\mathrm{S}} + r_{\mathrm{b,L}} \\ \mathbf{r}^{\mathrm{L}} &= \max(\mathbf{v}^{\mathrm{L}}, 0), \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} \tau_{\mathrm{S}}\dot{\mathbf{v}}^{\mathrm{S}} =& -(\mathbf{v}^{\mathrm{S}} - \mathbf{v}^{\mathrm{S}}_{\mathrm{leak}}) + \mathbf{A}^{\mathrm{u},+T}_{\mathrm{ON}}\mathbf{r}^{\mathrm{L}}_{\mathrm{ON}} + \mathbf{A}^{\mathrm{u},-T}_{\mathrm{ON}}\mathbf{r}^{\mathrm{L}}_{\mathrm{ON}} \\ &+ \mathbf{A}^{\mathrm{u},+\ T}_{\mathrm{OFF}}\mathbf{r}^{\mathrm{L}}_{\mathrm{OFF}} + \mathbf{A}^{\mathrm{u},-\ T}_{\mathrm{OFF}}\mathbf{r}^{\mathrm{L}}_{\mathrm{OFF}} + \mathbf{r}^{\mathrm{S}} \\ \mathbf{r}^{\mathrm{S}} =& \max(\mathbf{v}^{\mathrm{S}} - \lambda_{\mathrm{S}}, 0), \end{aligned} \tag{3.2}$$

where $\mathbf{x}^{\mathrm{L}} = [{\mathbf{x}^{\mathrm{L}}_{\mathrm{ON}}}^{T}, {\mathbf{x}^{\mathrm{L}}_{\mathrm{OFF}}}^{T}]^{T}$, $\mathbf{v}^{\mathrm{L}} = [{\mathbf{v}^{\mathrm{L}}_{\mathrm{ON}}}^{T}, {\mathbf{v}^{\mathrm{L}}_{\mathrm{OFF}}}^{T}]^{T}$, $\mathbf{r}^{\mathrm{L}} = [{\mathbf{r}^{\mathrm{L}}_{\mathrm{ON}}}^{T}, {\mathbf{r}^{\mathrm{L}}_{\mathrm{OFF}}}^{T}]^{T}$, $\mathbf{A}^{\mathrm{u},+} = [\mathbf{A}^{\mathrm{u},+}_{\mathrm{ON}}\ \mathbf{A}^{\mathrm{u},+}_{\mathrm{OFF}}]$,

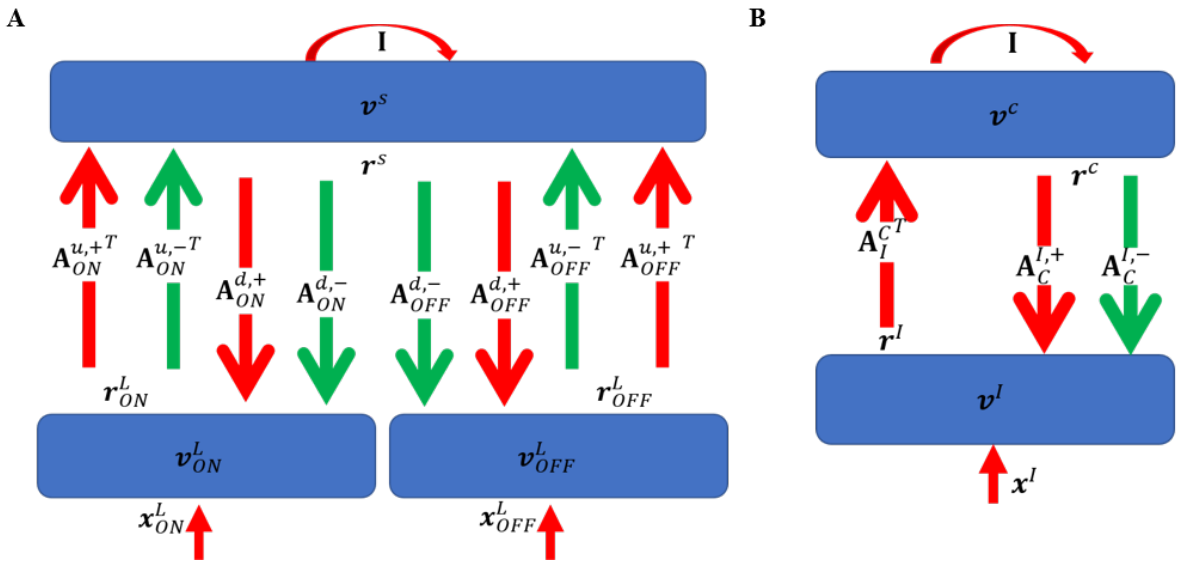

Figure 3.3: (A) Graphical representation of the bottom two-layer model. (B) Graphical representation of the top two-layer model. $\mathbf{I}$ is the identity matrix that represents self-excitation. Red and green arrows represent excitatory and inhibitory connections, respectively. Upward and downward arrows are for feedforward and feedback pathways. Notation defined in the main text.

$\mathbf{A}^{\mathrm{u},-} = [\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},-}\ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},-}]$, $\mathbf{A}^{\mathrm{d},+} = [\mathbf{A}_{\mathrm{ON}}^{\mathrm{d},+}\ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},+}]$, $\mathbf{A}^{\mathrm{d},-} = [\mathbf{A}_{\mathrm{ON}}^{\mathrm{d},-}\ \mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},-}]$, τ_{L} and τ_{S} are the time constants of the membranes of LGN cells and simple cells, $r_{\mathrm{b,L}}$ is the background firing rate for LGN cells, and λ_{S} is the threshold of the rectifying function of firing rates. $\mathbf{v}_{\mathrm{leak}}^{\mathrm{S}}$ represents the change of membrane potential caused by leakage currents. Details of the bottom two layers can be found in Chapter 2 (Lian et al., 2019)).

Similarly, the top two layers (graphical representation shown in Figure 3.3B) implement a second efficient coding model, but for complex cells receiving inputs from simple cells. The dynamics of the top two layers are given by

$$\begin{aligned} \tau_{\mathrm{I}}\dot{\mathbf{v}}^{\mathrm{I}} &= -\mathbf{v}^{\mathrm{I}} + \mathbf{x}^{\mathrm{I}} + \mathbf{A}_{\mathrm{C}}^{\mathrm{I}}\mathbf{r}^{\mathrm{C}} + r_{\mathrm{b,I}}\mathbb{1} \\ \mathbf{r}^{\mathrm{I}} &= \max(\mathbf{v}^{\mathrm{I}}, 0), \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} \tau_{\mathrm{C}}\dot{\mathbf{v}}^{\mathrm{C}} &= -\left(\mathbf{v}^{\mathrm{C}} - \mathbf{v}_{\mathrm{leak}}^{\mathrm{C}}\right) + \mathbf{A}_{\mathrm{I}}^{\mathrm{C}\,T}\mathbf{r}^{\mathrm{I}} + \mathbf{r}^{\mathrm{C}} \\ \mathbf{r}^{\mathrm{C}} &= \max(\mathbf{v}^{\mathrm{C}} - \lambda_{\mathrm{C}}, 0), \end{aligned} \tag{3.4}$$

where τ_{I}, $\mathbf{x}^{\mathrm{I}}$, $\mathbf{v}^{\mathrm{I}}$, $\mathbf{r}^{\mathrm{I}}$ and $r_{\mathrm{b,I}}$ are the time constant, input, membrane potential, firing rate, and background firing rate of intermediate cells in the third layer, $\mathbb{1}$ is a vector whose elements are all 1, τ_{C}, $\mathbf{v}^{\mathrm{C}}$, $\mathbf{r}^{\mathrm{C}}$ and λ_{C} are the time constant, membrane potentials, firing rates and firing threshold for complex cells in the fourth layer. $\mathbf{v}_{\mathrm{leak}}^{\mathrm{C}}$ represents the change of membrane potential for complex cells caused by leakage currents. $\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$ represents the

feedforward connection from intermediate cells to complex cells. Weights $\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$ are taken to be excitatory here so there are no inhibitory connections from intermediate cells to complex cells. The inputs to intermediate cells are the responses of simple cells so the excitatory connections between intermediate cells and complex cells indicate which simple cells are pooled for complex cells and how they are weighted. Introducing intermediate

Description	**Symbol**
Input stimuli to LGN/intermediate cells	$\mathbf{x}^{\mathrm{L}}$ / $\mathbf{x}^{\mathrm{I}}$
Input stimuli to ON/OFF LGN cells	$\mathbf{x}_{\mathrm{ON}}^{\mathrm{L}}$ / $\mathbf{x}_{\mathrm{OFF}}^{\mathrm{L}}$
Membrane time constant of LGN/simple/intermediate/complex cells (10 ms)	τ_{L} / τ_{S} / τ_{I} / τ_{C}
Membrane potentials of LGN/simple/intermediate/complex cells	$\mathbf{v}^{\mathrm{L}}$ / $\mathbf{v}^{\mathrm{S}}$ / $\mathbf{v}^{\mathrm{I}}$ / $\mathbf{v}^{\mathrm{C}}$
Membrane potentials of ON/OFF LGN cells	$\mathbf{v}_{\mathrm{ON}}^{\mathrm{L}}$ / $\mathbf{v}_{\mathrm{OFF}}^{\mathrm{L}}$
Firing rates of LGN/simple/intermediate/complex cells	$\mathbf{r}^{\mathrm{L}}$ / $\mathbf{r}^{\mathrm{S}}$ / $\mathbf{r}^{\mathrm{I}}$ / $\mathbf{r}^{\mathrm{C}}$
Firing rates of ON/OFF LGN cells	$\mathbf{r}_{\mathrm{ON}}^{\mathrm{L}}$ / $\mathbf{r}_{\mathrm{OFF}}^{\mathrm{L}}$
Spontaneous firing rate of LGN/intermediate cells (0.5 Hz)	$r_{\mathrm{b,L}}$ / $r_{\mathrm{b,I}}$
Leakage voltages of simple/complex cells	$\mathbf{v}_{\mathrm{leak}}^{\mathrm{S}}$ / $\mathbf{v}_{\mathrm{leak}}^{\mathrm{C}}$
Excitatory connection: all LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},+}$
Excitatory connection: ON/OFF LGN cells to simple cells	$\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},+}$ / $\mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},+}$
Inhibitory connection: all LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},-}$
Inhibitory connection: ON/OFF LGN cells to simple cells	$\mathbf{A}_{\mathrm{ON}}^{\mathrm{u},-}$ / $\mathbf{A}_{\mathrm{OFF}}^{\mathrm{u},-}$
Excitatory connection: simple cells to all LGN cells	$\mathbf{A}^{\mathrm{d},+}$
Excitatory connection: simple cells to ON/OFF LGN cells	$\mathbf{A}_{\mathrm{ON}}^{\mathrm{d},+}$ / $\mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},+}$
Inhibitory connection: simple cells to all LGN cells	$\mathbf{A}^{\mathrm{d},-}$
Inhibitory connection: simple cells to ON/OFF LGN cells	$\mathbf{A}_{\mathrm{ON}}^{\mathrm{d},-}$ / $\mathbf{A}_{\mathrm{OFF}}^{\mathrm{d},-}$
Excitatory connection: intermediate to complex cells	$\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$
Excitatory connection: complex to intermediate cells ($\mathbf{0}$)	$\mathbf{A}_{\mathrm{C}}^{\mathrm{I},+}$
Inhibitory connection: complex to intermediate cells	$\mathbf{A}_{\mathrm{C}}^{\mathrm{I},-}$
Sparsity level of simple / complex cells (both 0.1)	λ_{S} / λ_{C}
Upper bounds of LGN-simple/simple-complex connections (both 0.3)	$a_{1,max}$ / $a_{2,max}$
Learning rates of LGN-simple/simple-complex connections (3 and 0.5)	η_1 / η_2
Weight regulation constants of LGN-simple/simple-complex connections	γ_1 / γ_2

Table 3.1: Model symbols and parameters in Chapter 3.

cells separates the process of computing simple and complex cell responses so that we can simply investigate efficient coding for complex cells. $\mathbf{A}_{\mathrm{C}}^{\mathrm{I}}$ represents the feedback connection from complex cells to simple cells; i.e., $\mathbf{A}_{\mathrm{C}}^{\mathrm{I}} = \mathbf{A}_{\mathrm{C}}^{\mathrm{I},+} + \mathbf{A}_{\mathrm{C}}^{\mathrm{I},-}$, where $\mathbf{A}_{\mathrm{C}}^{\mathrm{I},+}$ are the excitatory connections from complex cells to simple cells and $\mathbf{A}_{\mathrm{C}}^{\mathrm{I},-}$ represent the inhibitory connections.

3.2.2 Input

We examine two different types of input to the model: static natural images and natural images with temporal information that simulates natural movement. Temporal information is prevalent in movements or videos of the real world, so it is important to consider this. The data set used in this chapter consists of 50 selected 1024×1536 pixel images of calibrated natural scenes from van Hateren's dataset (Van Hateren and Van Der Schaaf, 1998).

Static natural images

An $M \times M$ image patch chosen randomly from one selected 1024×1536 natural image is used as the input to the model. The responses of simple cells are then used as the inputs to the intermediate cells that feed into complex cells; i.e., $\mathbf{x}^{\mathrm{I}} = \mathbf{r}^{\mathrm{S}}$.

Natural images with jitter

For natural visual stimuli with temporal information, such as videos, the changing content between subsequent frames in one fixed region over a very short time period is very similar except for some translations or shifts in position, as it results from the movement of an object. Therefore, the temporal information in a natural video is similar to the temporal information in sequences of translated images, as used by Berkes and Wiskott (2005) to investigate temporal slowness. In this chapter, we simply use natural images with jitter to incorporate temporal information. The way of doing this is using sequences of random patches of an image region. The idea behind this is similar to it of using natural videos and sequences of translated images because images patches around the same location tend to have similar features except for translations or shifts.

We take N random image patches around a location to represent the image patches of temporal stimuli in one location over a short time period. More specifically, for a 1024×1536 natural image from the data set, a random location, (i_x, i_y), is chosen and then N image patches of size $M \times M$ whose centers are within the $0.4M$ pixel distance of (i_x, i_y) are randomly selected. For the chosen N image patches, the bottom two-layer model generates N sets of simple cell responses. Since the N image patches contain similar features with shifts (phase differences), the neural activities of simple cells in response to

the N image patches will also contain phase information. Using the concept of the trace rule where the response is determined by the current and past responses (Földiák, 1991), the average of N sets of simple cell responses is then used as the input to the complex cells; i.e., $\mathbf{x}^{\mathrm{I}} = \langle \mathbf{r}^{\mathrm{S}} \rangle$.

Since the image patches around the same location, (i_x, i_y), tend to have similar features but differ slightly in orientation, position, and size, the average responses of simple cells for these adjacent image patches integrate invariance and pass it to complex cells. Note that the input of static natural images is a special case ($N = 1$) in this scenario.

Pre-processing of natural stimuli

The input patches are first pre-whitened to mimic retinal processing before visual processing. This process is described as filtering the input image by ganglion cells whose RFs are characterized as divisively normalized difference-of-Gaussian filters (Tadmor and Tolhurst, 2000; Ratliff et al., 2010). The pre-whitened pixel intensity (I) at point (x, y) can be calculated by

$$I(x, y) = \frac{I_0(x, y) - I_1(x, y)}{I_d(x, y)} \tag{3.5}$$

where I_0, I_1, and I_d are responses measured by three unit-normalized Gaussian filters: center filter (g_0), surround filter (g_1), and divisive normalization filter (g_d), where g_d captures the local adaptation of ganglion cells (Troy et al., 1993). Images are convolved with g_0, g_1, and g_d. The standard deviation (SD) of the center filters is set to 1 pixel. The SD of surround filters is chosen to be 1.5 pixels or 1.5 times the center filter SD, which is consistent with previous measurements (Borghuis et al., 2008). The SD of g_d has the same size as g_1 (Ratliff et al., 2010).

The pre-whitened images are then multiplied with a 2D Gaussian windowing filter with SDs of 3 pixels in both vertical and horizontal directions. The purpose of this step is the same as the Gaussian synaptic distribution defined in Linsker's (1986) study and it puts more emphasis on the central part of the image patch to make the learned simple RFs more centralized in the 2D image domain. This step also assumes that model complex cells pool local simple cells that have RFs in the same region.

Pre-processed images are then fed into LGN cells in the first layer. If the pixel intensity is positive, it is set as the input to the ON LGN cell while the input to the corresponding OFF LGN cell is set to zero; if the pixel intensity is negative, the absolute value of the intensity is set as the input to the OFF LGN cell while the input to the corresponding ON LGN cell is set to zero.

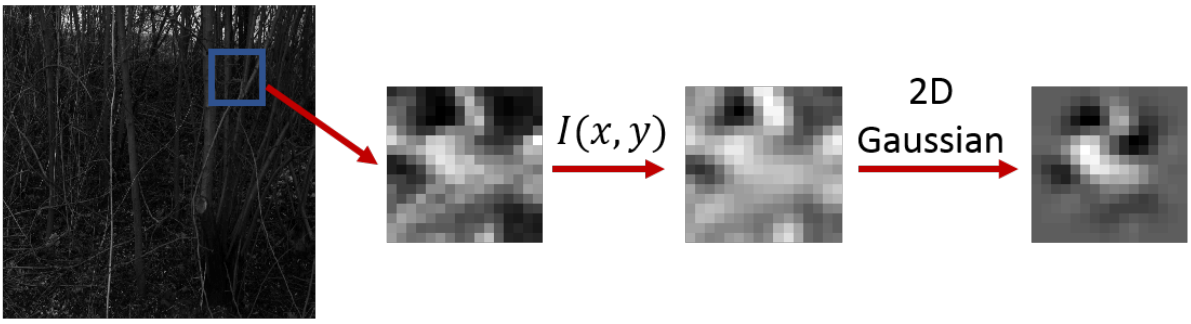

Figure 3.4: Pre-processing natural stimuli. The $M \times M$ image patch was pre-whitened by filters described by Eq. 3.5 and convolved with a 2D Gaussian filter to emphasize the central part of the image patch.

3.2.3 Learning rule

Learning LGN-simple cell connections

Connections between LGN and simple cells are learned based on Hebbian or anti-Hebbian plasticity; i.e., the changes of synaptic weights only depend on pre- and post-synaptic activities. The learning rule is from efficient coding and similar to (Chapter 2, Lian et al. (2019)), given by

$$\begin{aligned}
\Delta \mathbf{A}^{\mathrm{u},+} &= \eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{u},+} \right) \\
\Delta \mathbf{A}^{\mathrm{u},-} &= \eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{u},-} \right) \\
\Delta \mathbf{A}^{\mathrm{d},+} &= -\eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{d},+} \right) \\
\Delta \mathbf{A}^{\mathrm{d},-} &= -\eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{d},-} \right),
\end{aligned} \tag{3.6}$$

where η_1 is the learning rate, $\langle \cdot \rangle$ is the ensemble average operation over some samples, $\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}$ is the vector such that each element of vector $\mathbf{r}^{\mathrm{L}}$ is reduced by scalar $r_{\mathrm{b,L}}$, $(\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}})\mathbf{r}^{\mathrm{S}T}$ is the matrix given by the outer product of vectors $\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}$, $\mathbf{r}^{\mathrm{S}}$, and γ_1 is the weight regulation constant that prevents weights from growing without bound. $\mathbf{A}^{\mathrm{u},+}$ and $\mathbf{A}^{\mathrm{d},+}$ are kept non-negative while $\mathbf{A}^{\mathrm{u},-}$ and $\mathbf{A}^{\mathrm{d},-}$ are kept non-positive during learning. In addition, the absolute value of each weight is limited to an upper bound, $a_{1,\max}$, that represents the maximal synaptic efficacy. The only difference between the previous learning rules of simple cells (Chapter 2, Lian et al. (2019)) and this study is that weight normalization is replaced by the combination of self-regulation terms in Eq. 3.6 and the upper bound of connection weights.

Learning simple-complex cell connections

The top two layers implement a similar model to the bottom two layers. The connections between simple and complex cells are implemented via intermediate cells. The learning

rule comes from efficient coding and is similar to Eq. 3.6 and given by

$$\begin{aligned}
\Delta \mathbf{A}_{\mathrm{I}}^{\mathrm{C}} &= \eta_2 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,I}}) \mathbf{r}^{\mathrm{I}^T} \rangle - \gamma_2 \mathbf{A}_{\mathrm{I}}^{\mathrm{C}} \right) \\
\Delta \mathbf{A}_{\mathrm{C}}^{\mathrm{I},+} &= -\eta_2 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,I}}) \mathbf{r}^{\mathrm{I}^T} \rangle - \gamma_2 \mathbf{A}_{\mathrm{C}}^{\mathrm{I},+} \right) \\
\Delta \mathbf{A}_{\mathrm{C}}^{\mathrm{I},-} &= -\eta_2 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,I}}) \mathbf{r}^{\mathrm{I}^T} \rangle - \gamma_2 \mathbf{A}_{\mathrm{C}}^{\mathrm{I},-} \right),
\end{aligned} \tag{3.7}$$

where η_2 and γ_2 are the learning rate and weight regulation constant, respectively. In addition, the maximal synaptic weight allowed is $a_{2,\max}$.

3.2.4 Training

Input patches of size 16×16 ($M = 16$) are used in our model, similar to previous studies (Zylberberg et al., 2011; Zhu and Rozell, 2013; Lian et al., 2019), resulting in 256 ON and 256 OFF LGN cells. We use 100 simple cells and 100 intermediate cells in the second and third layers, respectively. For the number of complex cells in the fourth layer, two different values, 100 and 25, are used to investigate that whether the number of complex cells in the network is critical to learning complex cells. The time constants, τ_{L}, τ_{S}, τ_{I}, and τ_{C}, are taken to be 10 ms, which is physiologically plausible (Dayan and Abbott, 2001). The background firing rates, $r_{\mathrm{b,L}}$ and $r_{\mathrm{b,I}}$, are chosen to be 0.5 Hz; however, different values lead to similar results because the background firing rates only provide a working point for the dynamical model, as discussed in Section 2.5 (Lian et al., 2019). The dynamical system of the model described by Eq. 3.1-3.4 is numerically solved by the first-order Euler method. There are 20 iteration steps, with the integration time step of 4ms, for calculating the responses for both simple and complex cells. The models described in this chapter are implemented in MATLAB (version R2016b, MathWorks, MA, USA) using my own codes.

Simple cells

The bottom two layers of the network are trained first. Since during the course of the training $\mathbf{A}^{\mathrm{u},+}$ approaches $-\mathbf{A}^{\mathrm{d},-}$ and $\mathbf{A}^{\mathrm{u},-}$ approaches $-\mathbf{A}^{\mathrm{d},+}$, as described in Section 2.4 (Lian et al., 2019), we simply set $\mathbf{A}^{\mathrm{u},+} = -\mathbf{A}^{\mathrm{d},-}$ and $\mathbf{A}^{\mathrm{u},-} = -\mathbf{A}^{\mathrm{d},+}$ at the beginning of training. The upper bound of connection weights between LGN and simple cells, $a_{1,\max}$, is set to 0.3; i.e., the excitatory weights cannot exceed 0.3 and inhibitory weights cannot be less than -0.3. The sparsity level of simple cells, λ_{S} is set to 0.1. In addition, the learning rule (Eq. 3.6) used a batch that contained 100 randomly selected 16×16 image patches that have no temporal information in every epoch to accelerate the learning process. The weight regulation constant, γ_1, is set to 0.0001 and the learning rate, η_1, is 3. $100,000$ epochs are used in the training process. After the training process for simple cells, the connections between LGN and simple cells, $\mathbf{A}^{\mathrm{u}}$ and $\mathbf{A}^{\mathrm{d}}$, are fixed.

Complex cells

After the bottom two layers have been trained, the top two layers are then trained to learn complex cells. The natural images are used as the input to the first layer and simple cell responses generated in the second layer are the input to the top two layers where a rule based on efficient coding is used to learn the subspace of complex cells. Similarly, in accord with Section 2.4 (Lian et al., 2019), the feedforward excitatory (or inhibitory) connections converge to the opposite of the feedback inhibitory (or excitatory) connection, so it can be reasonably assumed that $\mathbf{A}_{\mathrm{C}}^{\mathrm{I},-} = -\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$ and $\mathbf{A}_{\mathrm{C}}^{\mathrm{I},+} = \mathbf{0}$ given that $\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$ is taken to be excitatory here. The maximal weight allowed for connections between simple and complex cells (via intermediate cells), $a_{1,\max}$, is 0.3. The learning rate, η_2, and the weight regulation constant, γ_2, are set to 0.5 and 0.0001, respectively. The sparsity level of complex cells, λ_{C}, is first set to 0.1 and then to 0 for a comparison. The learning process for complex cells runs for $100,000$ epochs.

Static natural images: The natural input is presented to the model and the responses of simple cells are then used as the input to intermediate cells; i.e., $\mathbf{x}^{\mathrm{I}} = \mathbf{r}^{\mathrm{S}}$. Similar to learning simple cells, a batch of 100 randomly selected image patches is used in every epoch.

Natural images with jitter: $N = 20$ random 16×16 image patches whose centers are within the $0.4M$ distance of (i_x, i_y) are chosen. For each of the 20 image patches, the model generates corresponding simple cell responses. The average of simple cell responses over 20 image patches is used as the input to intermediate cells; i.e., $\mathbf{x}^{\mathrm{I}} = \langle \mathbf{r}^{\mathrm{S}} \rangle$.

3.2.5 Spatial Phase invariance

Spatial Phase invariance, or partial invariance, is one of the most important features of complex cells. Here, sinusoidal gratings with different spatial phases are used as input to the trained model to examine whether model complex cells are invariant to different spatial phases.

Spatial phase tuning curve: First, an exhaustive search for each model complex cell was conducted to find the preferred orientation, spatial frequency, and spatial phase of the sinusoidal grating that evokes the maximal response in the following parameter space: orientation was varied between 0 and 180 degrees with steps of 15 degrees; spatial frequency was varied between 0.05 and 0.5 cycles/pixel with steps of 0.05 cycles/pixel; spatial phase was varied between 0 and 360 degrees with steps of 15 degrees. Then, a sequence of sinusoidal gratings was generated with the preferred orientation and spatial frequency and spatial phases spanning over 0 to 360 degrees with a step of 3.6 degrees (100 different spatial phases). This sequence of sinusoidal gratings is similar to the drifting sinusoidal gratings used in experimental studies. For each complex cell, the sequence of gratings with different spatial phases is used as the input to the model one after another,

while a sequence of responses for each grating is recorded. Therefore, a plot of responses vs. spatial phases can be plotted as the spatial phase tuning curve for each complex cell. A complex cell that is completely phase-invariant will have a flat spatial phase tuning curve, while a cell that is phase selective will have a bell-shaped spatial phase tuning curve.

F_1/F_0 ***ratio:*** Using cell activities in response to sinusoidal gratings, the ratio F_1/F_0 is used as a quantitative measure of spatial phase invariance (Skottun et al., 1991). F_1/F_0 ratio is defined as the ratio between the amplitude of Fourier Component of sinusoidal gratings at the drift rate and the amplitude of DC component of the gratings. In our case, the length of response and grating sequence is 100. The 1000-point FFT of the response sequence is first calculated, denoted as $R[k]$ where $k = 1,\ 2,\ \cdots$. The F_1/F_0 ratio can be computed by

$$F_1/F_0 = 2\frac{|R[100]|}{|R[0]|}, \tag{3.8}$$

where $R[0]$ and $R[100]$ are the 1st and 101st elements of sequence $R[k]$. The proof of Eq. 3.8 can be found in the Appendix A. A complex cell that is completely phase-invariant will have $F_1/F_0 = 0$, while cells that are selective to a small range of spatial phases will have F_1/F_0 ratios close to 2. Smaller values of F_1/F_0 indicate more spatial phase invariance.

3.2.6 The default linear model

After learning, the values of elements in the weight matrix $\mathbf{A}_{\mathrm{C}}^{\mathrm{I}}$ indicate how simple cells are pooled by complex cells. In order to investigate the subspace pooled by each model complex cell, a control model of complex cells that linearly sums over the responses of pooled simple cells using learned connections is used, as given by

$$\mathbf{r}^{\mathrm{C}} = \mathbf{A}_{\mathrm{I}}^{\mathrm{C}^T}\mathbf{r}^{\mathrm{S}}. \tag{3.9}$$

The control model will be called the default linear model throughout this chapter.

The default linear model uses the connection weights learned by our efficient coding model, but it uses a different method, linear summation, of computing complex cell responses compared with the efficient coding model. Two important questions that arise here are (1) does the subspace contain simple cells that are similar to those observed experimentally and that have similar orientations but different spatial phases and (2) does the principle of efficient coding reduce spatial phase invariance for complex cells?

3.3 Results

After training the connection weights in the first two layers, similar to Section 2.4 (Lian et al., 2019), we used the *synaptic field* ($\mathbf{S}_\mathrm{f}$) defined as

$$\mathbf{S}_\mathrm{f} = (\mathbf{A}_\mathrm{ON}^{\mathrm{u},+} + \mathbf{A}_\mathrm{ON}^{\mathrm{u},-}) - (\mathbf{A}_\mathrm{OFF}^{\mathrm{u},+} + \mathbf{A}_\mathrm{OFF}^{\mathrm{u},-}) \tag{3.10}$$

to visualize the overall effect of all connections between ON and OFF LGN cells and simple cells. Figure 3.5 shows the synaptic fields for 100 model simple cells. Compared with synaptic fields in Figure 2.5, the $\mathbf{S}_\mathrm{f}$ here were more centralized in the 16×16 image region, as a consequence of introducing a 2D Gaussian window (illustrated in Figure 3.4) to emphasize the central part of input images.

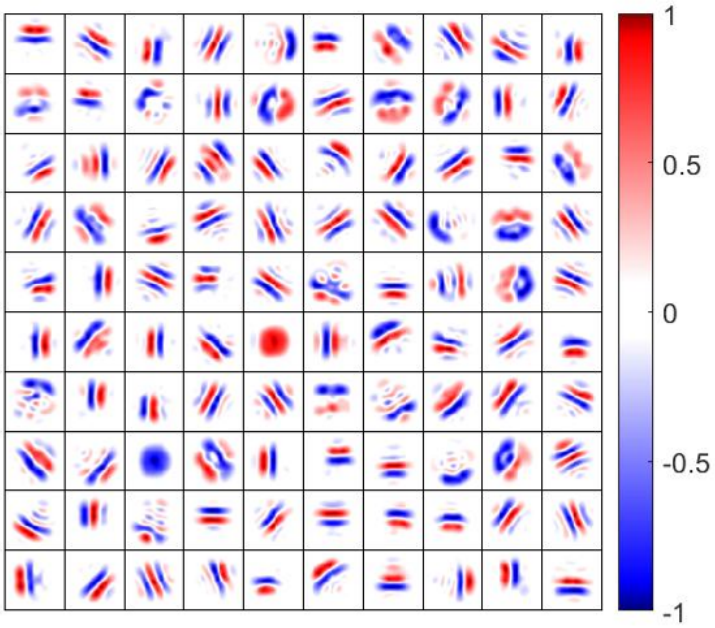

Figure 3.5: Synaptic fields (defined in Eq. 3.10) for 100 simple cells. Each block is a 16×16 image that displays the overall effects of LGN cells and simple cells. 100 cells are located on a 10×10 grid. Values in each block are normalized to the range $[-1\ 1]$.

3.3.1 Efficient coding trained on static natural images fails to pool simple cells to form the subspace of complex cells

Hyvärinen and Hoyer (2000) applied efficient coding (ICA) to find the subspace that is able to generate complex cell properties. However, their model has two important assumptions that are unrealistic for the biological neural network: (1) a two-sided power nonlinearity and (2) fixed weights between the model complex cell and the units in its subspace. We removed these two assumptions in our model and trained the model on static natural images. Our simulation results show that efficient coding for complex cell

cannot pool simple cells to form the subspace of complex cells who are phase invariant, if the model incorporates more biologically constraints and is trained on static natural images that have no temporal information.

After training the model with $\lambda_c = 0.1$ and 100 complex cells, many weights of the connection between simple and complex cells reduce to small values around zero, while some weights are significant. Each weight of feedforward connection, $\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$, represents how strongly a simple cell is pooled by a complex cell.

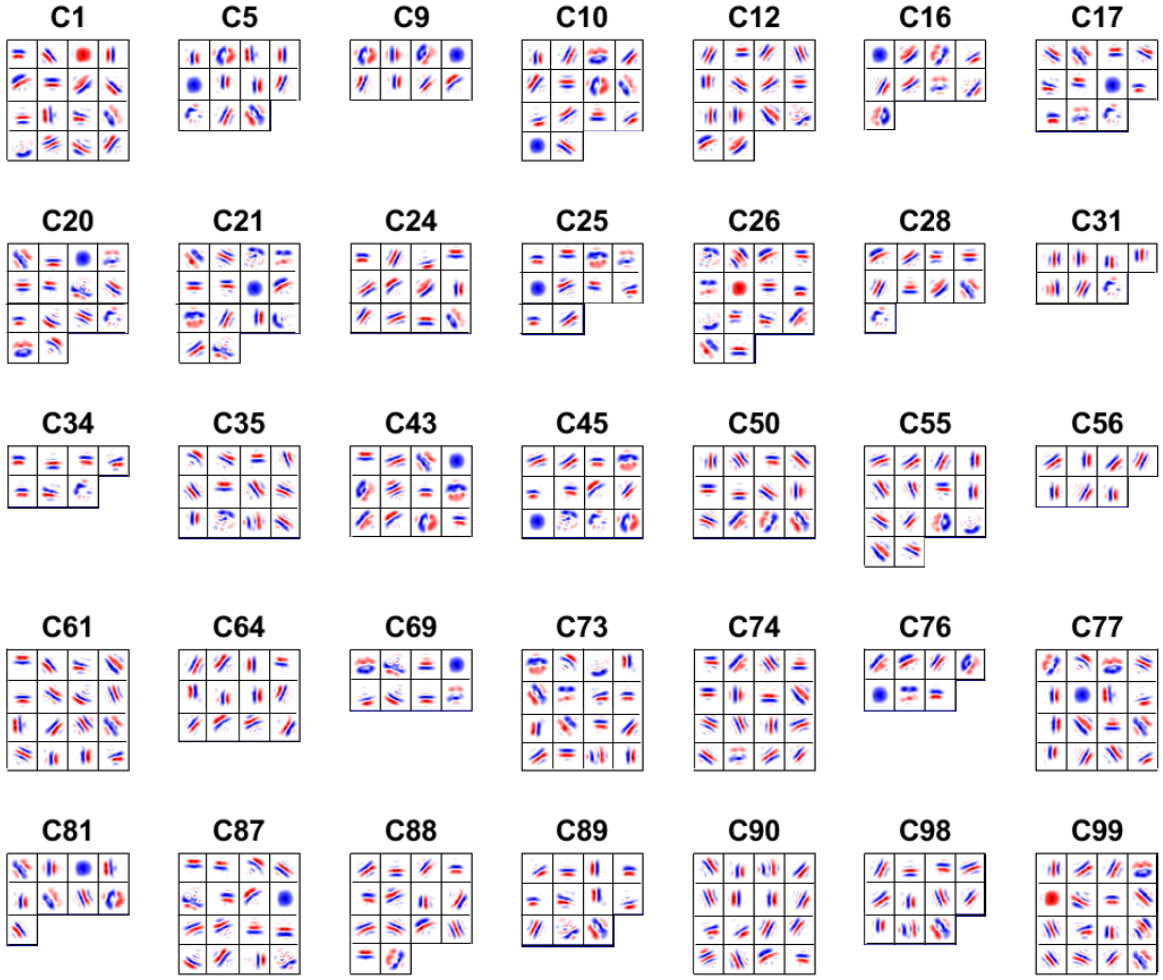

Figure 3.6: Complex cell subspace with $\lambda_c = 0.1$ and 100 complex cells. C represents complex cell and the followed number is the index of the complex cell. Each block is a 16×16 synaptic field (defined in Eq. 3.10) of simple cells that have a feedforward weight larger than 0.1 in the subspace of a complex cell. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. The figure only shows up to 16 subunits.

Figure 3.6 shows the subspace of 35 randomly selected complex cells (a full subspace of model complex cells can be seen in Figure B.1 of Appendix B). Each block in any subspace of the complex cell represents the 16×16 synaptic field of a simple cell that has a feedforward weight larger than 0.1, which means that only significant simple cells are displayed here. The figure shows that most complex cells have more than 8 simple cells in

the subspace. The average number of simple cells in the subspace among all complex cells is 12.3 (with standard deviation 3.27), indicating that each complex cell normally pools many simple cells.

Furthermore, most model complex cells in Figure 3.6 pool simple cells with a wide range of orientations, suggesting that the learned subspaces are not selective to orientations, which contradicts with orientation tuning of complex cells. We next show the spatial phase tuning curves for a few examples of model complex cells.

The complex cell that pools simple cells with similar orientations Complex cell C34 (Figure 3.6 and 3.7) has 7 simple cells in the subspace. Apart from the last simple cell in Figure 3.7A, pooled simple cells have similar orientations. However, this type of model complex cells is very rare, as can be seen from the subspaces shown in Figure 3.6. Figure 3.7B displays the spatial phase tuning curve of complex cell C34 and shows an interesting fact that spatial phase tuning curves of simple cells are largely overlapped and only cover a limited region of the spatial phase. As a result, the model complex cell is only invariant to a small region of spatial phase. A large value of $F_1/F_0 = 1.76$ also shows that the model complex cells C34 is actually simple-cell like. The spatial phase tuning curve for the default linear model is similar to C34 except the curve is shifted above due to simple cell S13. The shift of tuning curve is not observed for C34, because efficient coding model brings competition and the responses are suppressed compared to a linear model (Olshausen and Field, 1997).

The complex cell that pools simple cells with various orientations Complex cell C74 (Figure 3.6 and 3.8) has 16 significant simple cells (weights larger than 0.1) in the subspace and pooled simple cells have many orientations: horizontal, vertical and diagonal orientations, suggesting that the model fails to selectively pool simple cells with similar orientations. This kind of model complex cells is a majority, as seen from the subspaces shown in Figure 3.6. The spatial phase tuning curve of complex cell C74 (Figure 3.8B) shows two interesting phenomena: (1) for a preferred sinusoidal gratings with a fixed orientation, simple cells with different orientations contribute to spatial phase invariance for that orientation, but the contribution is little (seen from the relatively small amplitudes of tuning curves); (2) the tuning curves of the default linear model is very different from model cell C74, where default linear model covers a much wider region of spatial phase than the efficient coding model.

Problems of the current model

Above all, the current efficient coding model for complex cells trained on natural images have several problems. First, pooled simple cells in the subspace show almost no similarity in orientation, which contradicts with orientation selectivity of complex cells; in other words, the current model fails to pool simple cells with similar orientations into the subspace of

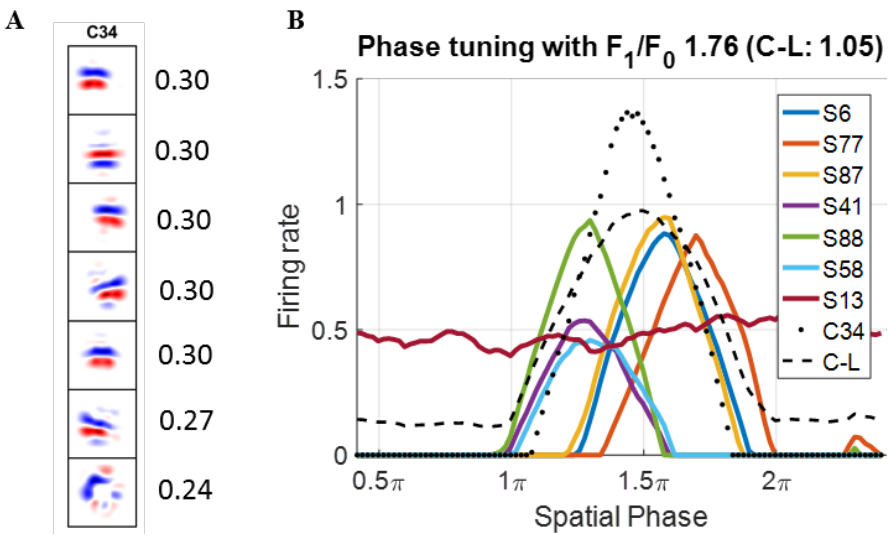

Figure 3.7: Complex cell C34. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C34. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell C34. S represents simple cell and the following number is the index of the simple cell. C-L represents the control model of complex cells that linearly sums simple cell responses weighted by the connection weights.

complex cells. Second, the principle of efficient coding suppresses cell responses compared with the default linear model.

For the first problem, one reason is that model complex cells tend to be inactive; i.e., in response to a stimulus, only a small portion of complex cells will have non-zero responses. In our simulation, the percentage of model complex cells having non-zero responses during training is 8.1%, which means on average only a few complex cells are trying to represent simple cell activities in response to each stimulus. As a consequence, those complex cells must pool many different simple cells to account for the diversity of simple cell responses, which means that complex cells fail to selectively pool simple cells. According to the model dynamics, the sparsity level of complex cells, λ_C, controls the sparseness of complex cells. Therefore, reducing λ_C will increase the percentage of complex cells being active and encourage complex cells to pool simple cells selectively.

For the second problem, it is also related to the choice of λ_C. Larger λ_C indicates that a smaller percentage of complex cells will be active, so the complex cell responses will be largely suppressed.

Therefore, combining the causes of the two problems mentioned above, it is natural to reduce λ_C in order to solve the issues of the current model. In the next section, results of model complex cells with λ_C will be analyzed.

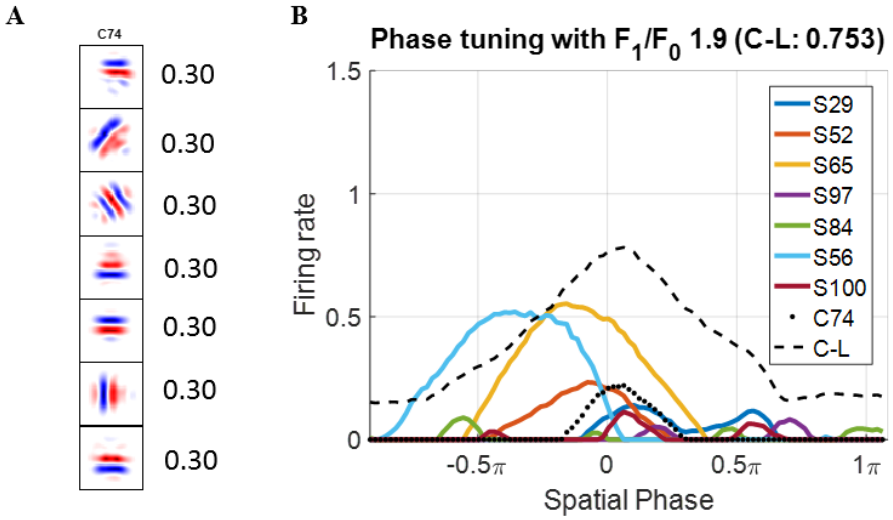

Figure 3.8: Complex cell C74. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C74. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell C74. S represents simple cell and the following number is the index of the simple cell. C-L represents the default linear model of the complex cell that linearly sums simple cell responses weighted by the connection weights. Up to 7 simple cells in the subspace are displayed.

Setting sparsity level to zero

The sparsity level of complex cells, λ_{C}, is the firing threshold and the minimum value allowed is zero. In this section, results were generated using the same model except that λ_{C} was set to zero. We show that even with the smallest possible value of λ_{C}, efficient coding for complex cells trained on static natural images still fails to account for the spatial phase invariance of complex cells.

Learned subspace: After learning, most connection weights between simple and complex cells reduce to very small values around zero, with only a few significant weights. Each weight of feedforward connection, $\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$, represents how strongly a simple cell is pooled by a complex cell. Figure 3.9 shows the subspace of 35 randomly selected complex cells (a full subspace of model complex cells can be seen in Figure B.2 of Appendix B). Most complex cells have two or three simple cells in the subspace. The average number of simple cells in the subspace among all complex cells is 2.16 with the standard deviation of 0.60. Each block in any subspace of the complex cell represents the 16×16 synaptic field of a simple cell that has a feedforward weight larger than 0.1, which means that only significant simple cells are displayed here.

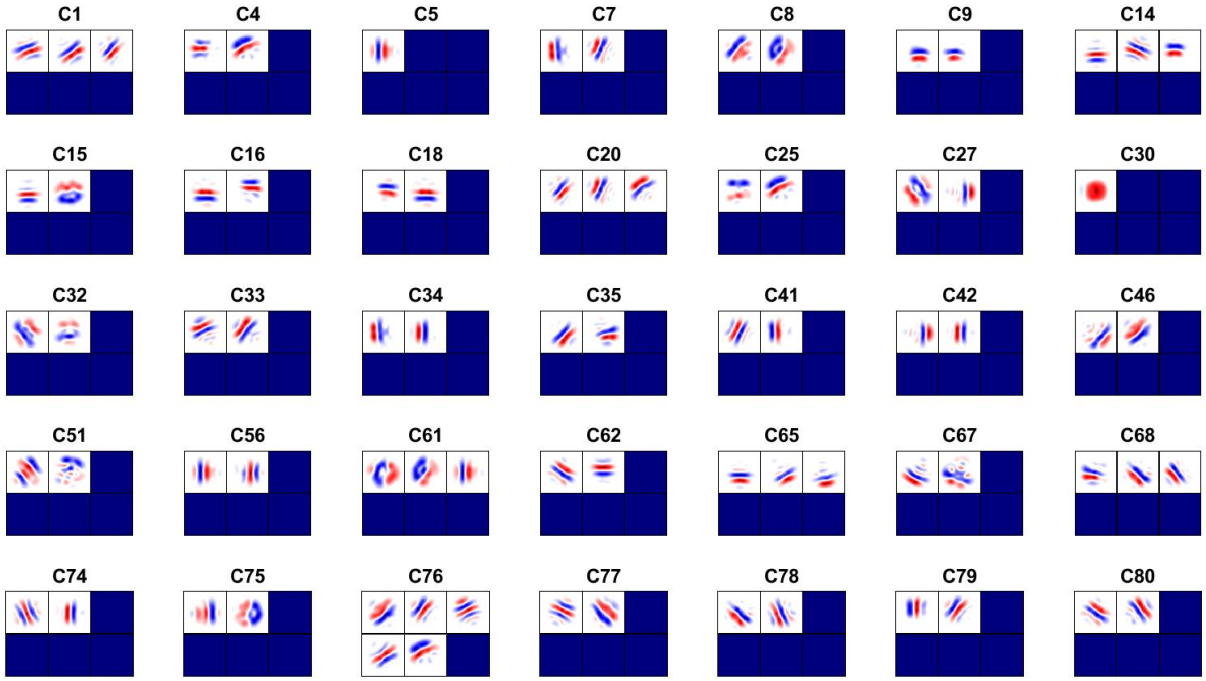

Figure 3.9: Complex cell subspace with $\lambda_c = 0$ and 100 complex cells. C represents complex cell and the followed number is the index of the complex cell. Each block is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace of a complex cell. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure.

Compared with the model with $\lambda_{\mathrm{C}} = 0.1$ in the previous section, simple cells in the learned subspace are less and have similar orientations (at most two distinct orientations), indicating that the model can now selectively pool simple cells. Furthermore, the

percentage of model complex cells being active is now 45.4%.

As seen in Figure 3.9, there is a wide range of subspaces for complex cells. For example, complex cell C5 and complex cell C30 only have one significant simple cell in the subspace, while other complex cells have two or more. In addition, the simple cells in the subspace of some complex cells have similar orientation, such as complex cell C18 and complex cell C42, while the simple cells in the subspace of some other complex cells have distinct orientations, such as complex cell C4 and complex cell C74.

Some example complex cells are examined in detail to show that the learned subspace of complex cells using static natural images is insufficient to explain spatial phase invariance of complex cells. These examples are chosen because they are representative of the diversity of model complex cells with more than two subunits.

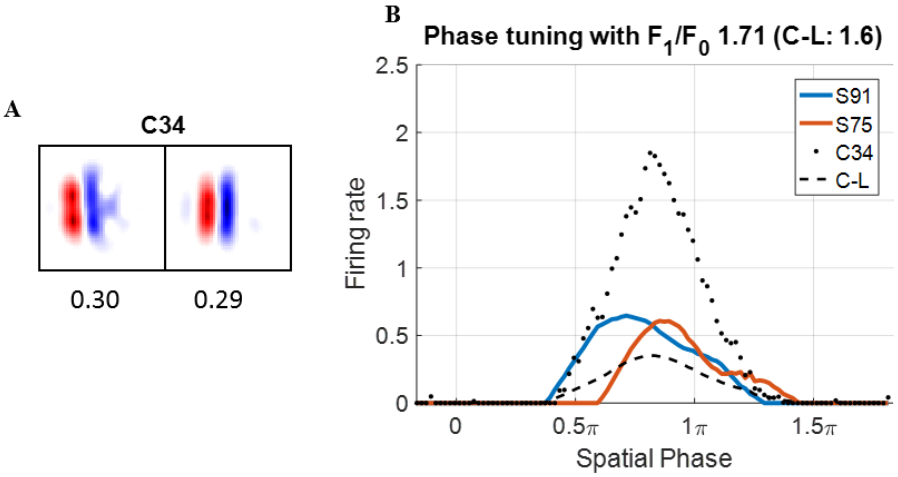

Figure 3.10: Complex cell C34. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number below is the feedforward weight connected with complex cell C34. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the control model of complex cells that linearly sums simple cell responses weighted by the connection weights.

The complex cell Pools simple cells that have the same orientation but different spatial phase preferences: Figure 3.10 shows complex cell C34 where the simple cells in the subspace have similar orientations but around $\pi/2$ radians spatial phase difference. By pooling simple cells with similar orientations but different preferred spatial phases, complex cell C34 becomes more invariant to spatial phase. However, the diversity in spatial phase tuning of simple cells in the subspace is limited and insufficient to generate spatial phase invariance for the complex cell. $F_1/F_0 = 1.71$ indicates that the model complex cell is highly selective to phase and behaves like real simple cells. The

tuning curve of the default linear model is similar to our model with a large value of F_1/F_0 (1.6).

The complex cell pools simple cells that have similar orientation and spatial phase tuning but differ in positions: Figure 3.11 shows an example of complex cell, C9, whose subspace has simple cells with similar orientation tuning and spatial phase tuning. There are two simple cells in the subspace, but the first synaptic field is slightly shifted to the right compared with the second one. Figure 3.11B shows that the two simple cells have similar spatial phase tuning with a preferred spatial phase at around 2π radians. Therefore, the complex cell has similar spatial phase preference and does not display spatial phase invariance, which can also be observed from the large value of F_1/F_0 (1.83). The F_1/F_0 ratio for the default linear model is also very large (1.64), indicating that the simple cells in the subspace are not sufficient to produce spatial phase invariance for this complex cell.

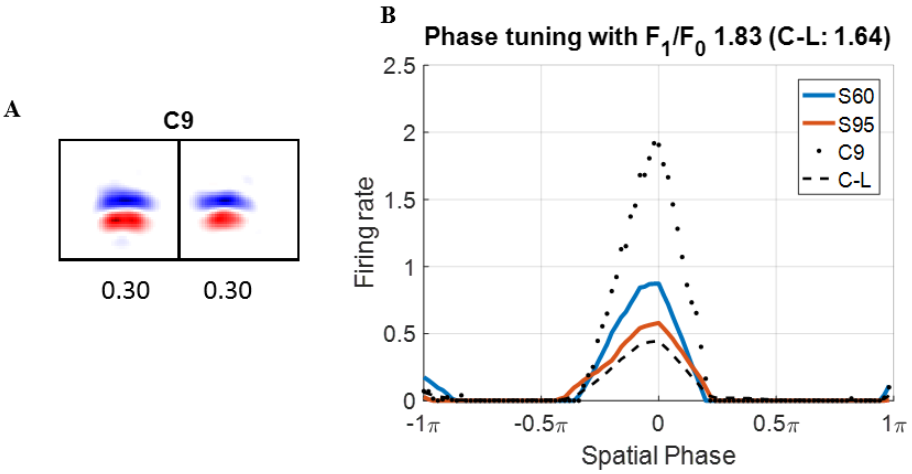

Figure 3.11: Complex cell C9. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number below is the feedforward weight connected with complex cell C9. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell C9. S represents simple cell and the following number is the index of the simple cell. C-L represents the control model of complex cells that linearly sums simple cell responses weighted by the connection weights.

The complex cell pools simple cells that have distinct orientation preferences: Figure 3.12 shows another example of complex cell, C74, that has simple cells with distinct orientations in the subspace of the complex cell. The simple cell with distinct orientation in the subspace can also contribute to spatial phase invariance for the complex cell. Though simple cells in the subspace have different spatial phase preferences, the tuning curve of model complex cell C74 is narrower than the union of tuning curves of simple cells in the subspace, which can also be seen from a larger F_1/F_0 ratio (1.67) of

the C74 compared with the F_1/F_0 rato of default linear model (1.42).

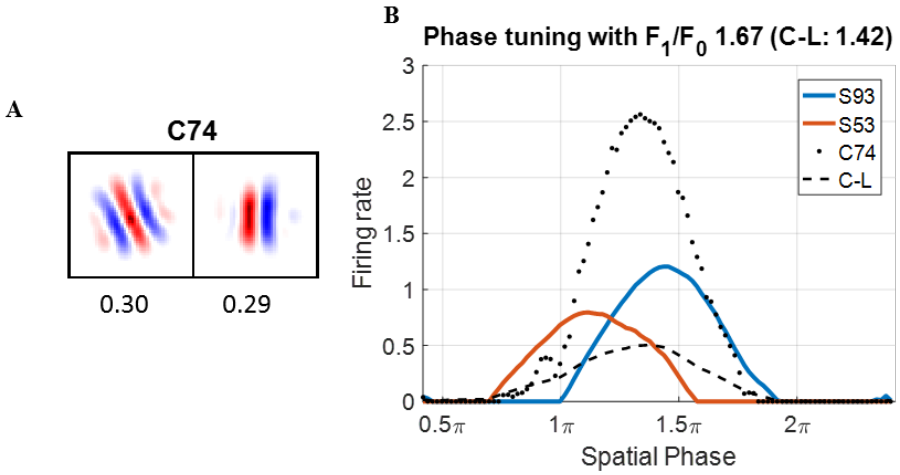

Figure 3.12: Complex cell C74. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number below is the feedforward weight connected with complex cell C74. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the control model of complex cells that linearly sums simple cell responses weighted by the connection weights.

Pooled simple cells that have similar orientations: Figure 3.13 shows complex cell C1 with all three simple cells in the subspace with similar orientations, which brings more spatial phase invariance to the complex cell. However, the complex cell is still highly selective to spatial phase with a large value of the F_1/F_0 ratio (1.7). The F_1/F_0 ratio (1.53) of the default linear model is smaller than model complex cell, but it is still large and indicates limited spatial phase invariance.

After examining some examples of model complex cells, distributions of F_1/F_0 ratio for experimental complex cells, model complex cells, and the default linear models will be shown next to illustrate that current model fails to explain spatial phase invariance of complex cells.

Histogram of F_1/F_0 implies that the learned subspace is not able to produce spatial phase invariance for complex cells found in the experiment: Figure 3.14 shows that the histograms of F_1/F_0 have similar distributions of F_1/F_0 ratio for experimental complex cells, model complex cells and the default linear model. For the experimental complex cells, cells with $F_1/F_0 < 1$ in the study (Ringach et al., 2002) are included because a generally accepted classification of complex cells is that they have and F_1/F_0 ratio smaller than 1 (Skottun et al., 1991). Figure 3.14A shows that the distribution of F_0/F_1 ratio for experimental complex cells spreads over the interval $[0\ 1]$ but skews toward zero. However, the distribution of F_1/F_0 ratios for model complex cells

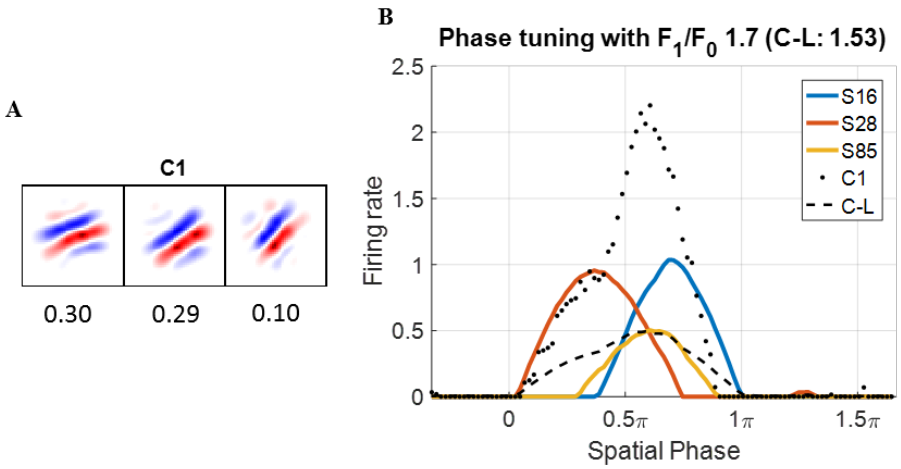

Figure 3.13: Complex cell C1. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number below is the feedforward weight connected with complex cell C1. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the default model of complex cells that linearly sums simple cell responses weighted by the connection weights.

(Figure 3.14B) is centered around 1.5 and only covers the interval [1 2], suggesting that most model complex cells are not spatial phase invariant like simple cells. The distribution for the default linear model (Figure 3.14C) is similar to model complex cells, suggesting that the simple cells in the learned subspace are not sufficient to provide spatial phase invariance for complex cells.

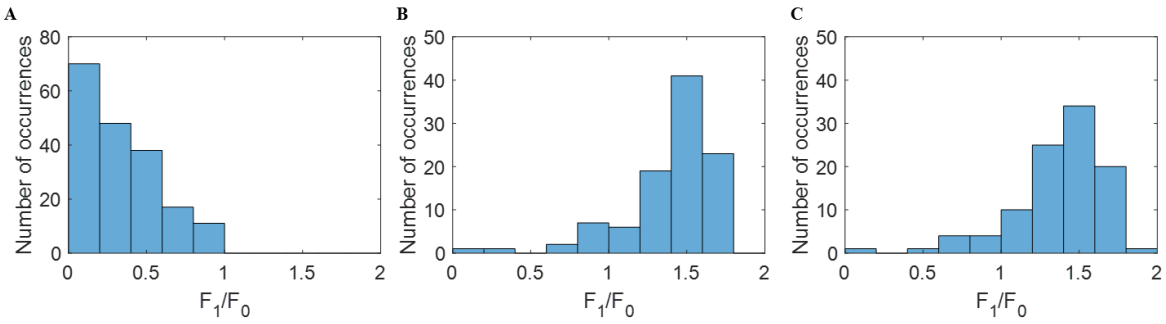

Figure 3.14: Histograms of F_1/F_0 with $\lambda_c = 0$ and 100 complex cells. (A) Experimental complex cells (Ringach et al., 2002). (B) Model complex cells. (C) default linear model of complex cells.

Problems of the current model: The above results indicate that our efficient coding model with $\lambda_C = 0$ can pool simple cells with similar orientation but different spatial phases when the model is trained on simple cell responses using static natural images. Because of the reduced value of λ_C, the two problems of our model with $\lambda_C = 0.1$

are resolved: (1) learned complex cells can pool simple cells with selective orientations; (2) responses of model complex cells are not severely suppressed compared with the default linear model (seen from a similar distribution of F_1/F_0 ratios in Figure 3.14B and C).

However, there is still a problem with the current model: pooled simple cells are not sufficient to cover a wider region of spatial phase. The squaring nonlinearity in ISA naturally introduces spatial phase with π difference (Hyvärinen and Hoyer, 2000), but our model will need to pool simple cells with various spatial phase preferences to form the subspace of complex cells such that the total region of spatial phase contributed by simple cells could be much larger.

To solve this problem, each model complex cell should pool more simple cells with similar orientations but different spatial phase preferences. In previous studies of modeling complex cells (Einhäuser et al., 2002; Hyvärinen and Hoyer, 2000), the number of model complex cells is much smaller than the number of simple cells. Especially for the study of Einhäuser et al., the numbers of complex cells and simple cells are 60 and 4 so that each complex cell can pool many simple cells. Therefore, we further change the number of complex cells to 25 instead of 100 to investigate whether our model of efficient coding for complex cells can account for spatial phase invariance after learning. The results are analyzed in the next section.

Using less model complex cells

If we set the number of model complex cells to 25 (100 model simple cells) and trained the model on static natural images, following results demonstrate that dramatically reducing the number of model complex cells does not help the model learn spatial phase invariant complex cells.

Learned subspaces: After learning, most connection weights between simple and complex cells reduce to very small values around zero, with only a few significant weights. Each weight of feedforward connection, $\mathbf{A}_{\mathrm{I}}^{\mathrm{C}}$, represents how strongly a simple cell is pooled by a complex cell. Figure 3.9 shows the subspace of all 25 model complex cells. Each block in any subspace of the complex cell represents the 16×16 synaptic field of a simple cell that has a feedforward weight larger than 0.1, which means that only significant simple cells are displayed here. Most complex cells have four or more simple cells in the subspace. The average number of simple cells in the subspace among all complex cells is 5.52 with the standard deviation of 1.71. Compared with the model that has 100 complex cells in the previous section, the current model has pooled more simple cells in the subspace. However, we will next show that this does not make the model learn spatial phase invariance of complex cells.

One complex cell that pools simple cells with the same orientation and similar spatial phase preferences: Figure 3.16A shows a complex cell that pools 5

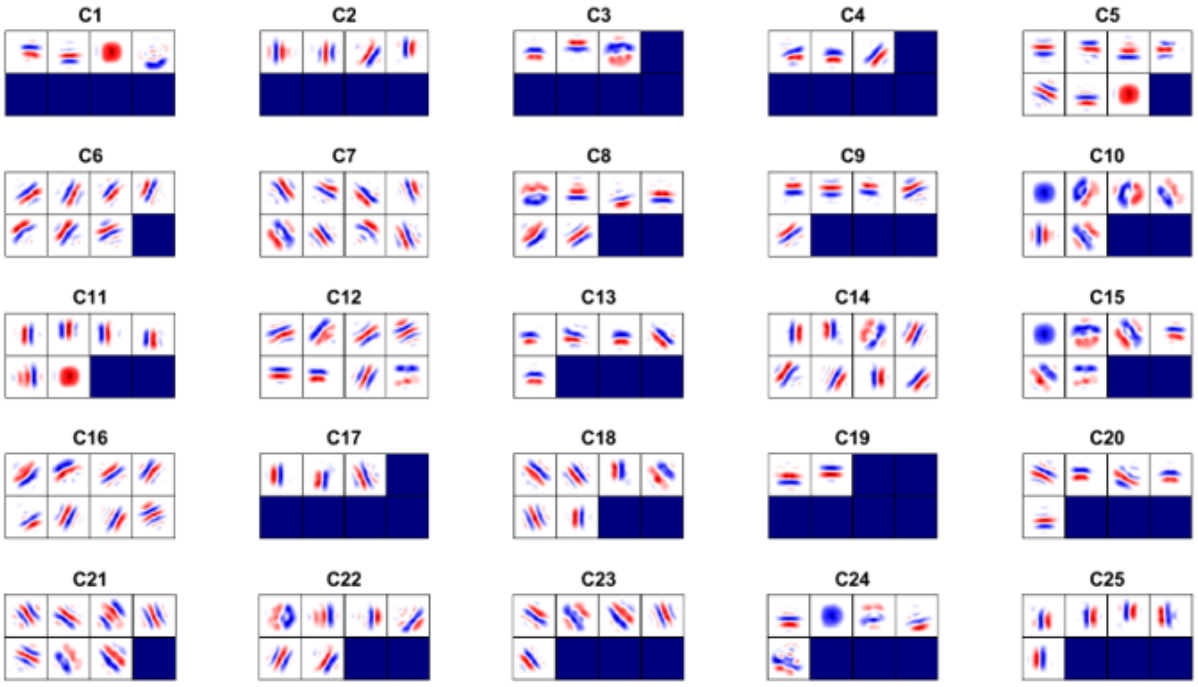

Figure 3.15: Subspace of all 25 model complex cells with $\lambda_c = 0$ and 25 complex cells. C represents complex cell and the followed number is the index of the complex cell. Each block is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace of a complex cell. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure.

significant simple cells with the same orientation. However, pooled simple cells have very similar spatial phase preferences so that the model complex cell is highly selective to one spatial phase and displays no invariance, which can also be seen from a large F_1/F_0 ratio (1.68).

One complex cell pools simple cells with the same orientation but various spatial phase preferences: Figure 3.17 shows a complex cell that pools 6 significant simple cells with the same orientation except for the last simple cell. The spatial phase tuning curves in Figure 3.17B demonstrate that simple cells with the vertical orientation nearly cover the whole region of the spatial phase. Therefore, the F_1/F_0 ratio of the default linear model (0.698) is smaller than one, indicating that simple cells in the subspace are sufficient to produce spatial phase invariance for the complex cell. However, F_1/F_0 ratio of the model complex cell is still larger than 1, suggesting that the response of the model complex cell is suppressed so that the model complex cell becomes less invariant to spatial phase. However, this type of model complex cells is rare as can be seen from the distribution of F_1/F_0 shown below (Figure 3.18B and C).

Histogram of F_1/F_0 implies that using less model complex cells does not help the model learn spatial phase invariance of complex cells found in the experiment: Figure 3.18 shows the distributions of F_1/F_0 for experimental complex cells, model complex cells, and the default linear model. Seen from Figure 3.18C, the distribution

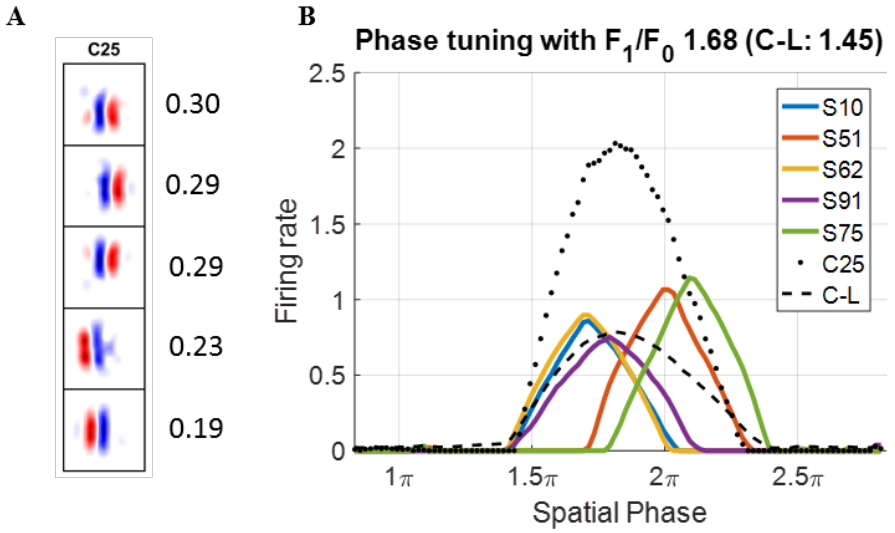

Figure 3.16: Complex cell C25. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C25. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the default model of complex cells that linearly sums simple cell responses weighted by the connection weights.

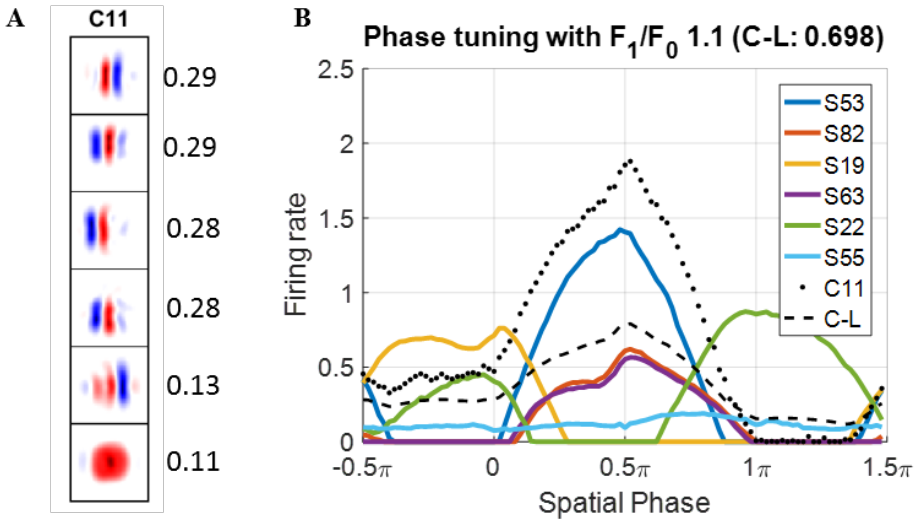

Figure 3.17: Complex cell C11. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C11. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the default model of complex cells that linearly sums simple cell responses weighted by the connection weights.

of F_1/F_0 for the default linear model has slightly shifted to low values compared with Figure 3.14C, indicating that having less complex cells helps model complex cells pool simple cells with various spatial phase preferences, though default linear model of complex cells with $F_1/F_0 < 1$ is only a minority. Figure 3.14B shows that the distribution of F_1/F_0 for model complex cells is still highly skewed towards 2 (shifted to the right of the distribution for the default liner model), which indicates that the efficient coding model of complex cells slightly suppresses model activities compared with a linear model.

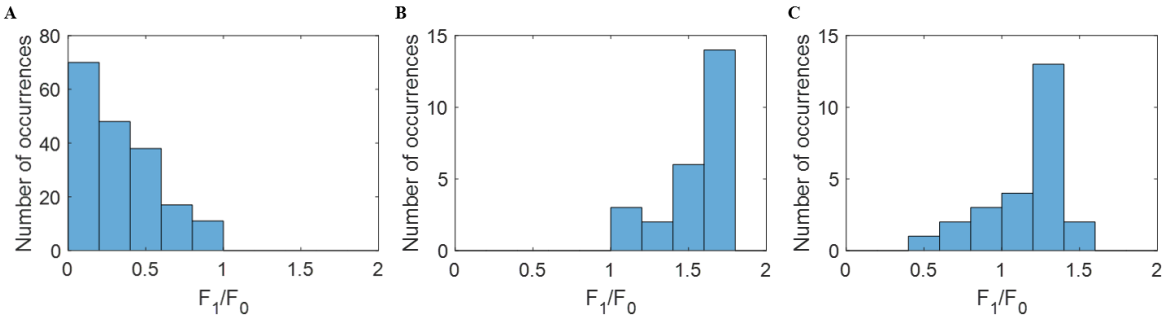

Figure 3.18: Histograms of F_1/F_0 with $\lambda_c = 0$ and 25 complex cells. (A) Experimental complex cells (Ringach et al., 2002). (B) Model complex cells. (C) default linear model of complex cells.

Problems with the current model: Having 25 model complex cells and 100 simple cells is not biologically plausible, because complex cells are the most common cell in V1 (Hubel and Wiesel, 1968). Though reducing the number of model complex cells to one-fourth of the simple cells, the efficient coding model of complex cells trained on static natural images fails to learn the property of spatial phase invariance and most model complex cells should be categorized as simple cells because of large values of F_1/F_0 ratio (Figure 3.14B). The current model can pool simple cells with more diverse spatial phase preferences that contribute to a small F_1/F_0 for the default linear model. However, the distribution of F_1/F_0 for the default linear model is dominant by values larger than 1. Furthermore, the fact that the distribution of F_1/F_0 for the efficient coding model is shifted right compared with the default linear model suggests the current model suppresses model cell activities even though the sparsity level of complex cells, λ_C, is set to zero.

After smaller values of λ_C and number of model complex cells are used, efficient coding model of complex cells trained on static natural images still fails to pool simple cells with diverse spatial phase preferences such that the union of spatial phase tuning can account for spatial phase invariance.

Next, the results of the model trained on natural images with temporal information will be analyzed.

3.3.2 Efficient coding for complex cells trained on natural images with jitter fails to explain complex cells properties

Based on the results in previous sections, a model with the sparsity level of complex cells, λ_C, setting zero, and 100 model complex cells (the same as simple cells) is trained on natural images with jitter. Simulation results show that efficient coding model of complex cells can pool simple cells with similar orientations but a wide range of spatial phase preferences, but the network dynamics of efficient coding suppresses model cell responses so that spatial phase invariance cannot be generated.

After training the model using natural images with jitter, most connection weights reduce to small values while only a few weights are significant, i.e., with values larger than 0.1. Similar to Figure 3.9, the subspaces of 35 randomly selected complex cells trained on images with temporal information are displayed in Figure 3.19 (full subspaces can be found in Figure B.3 of Appendix B). Figure 3.19 shows that most complex cells have more than 5 simple cells (average number of simple cells in the subspace among all complex cells: 6.68). Figure 3.19 shows a wide range of subspaces: some have simple cells with the same orientations (e.g., complex cell C31); some have simple cells with two distinct orientations (e.g., complex cell C22); and some have simple cells with similar but different orientations (e.g., complex cell C5).

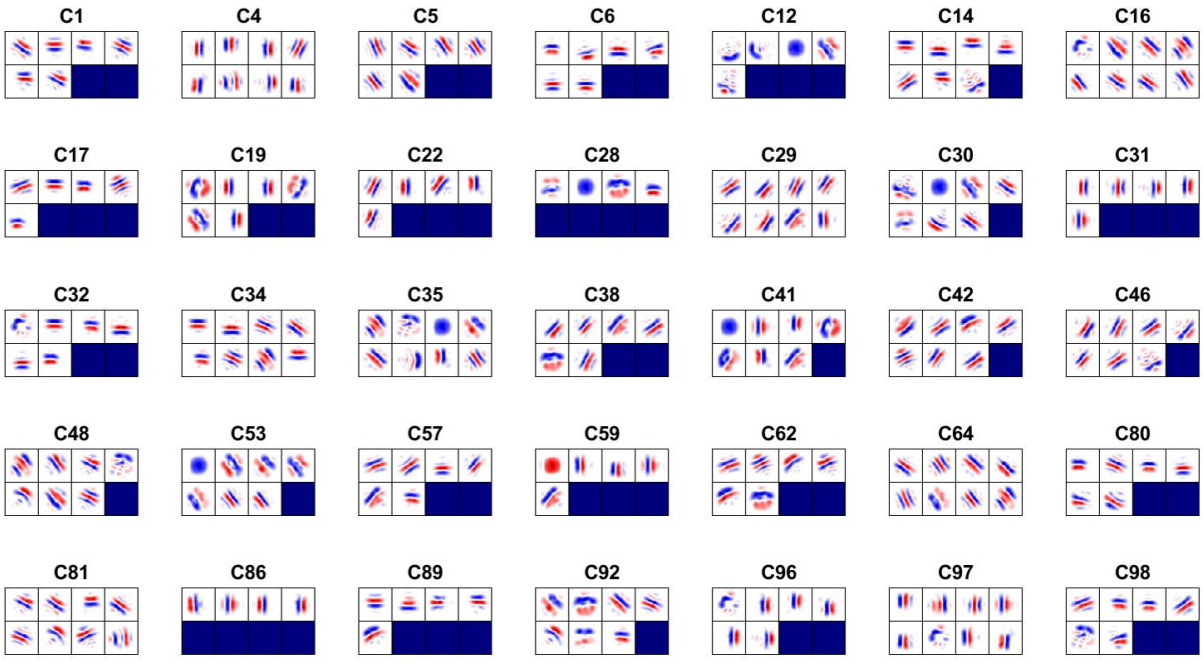

Figure 3.19: Complex cell subspace with $\lambda_c = 0$ and 100 complex cells trained on natural images with jitter. C represents complex cell and the following number is the index of the complex cell. Each block is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace of a complex cell. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure.

Examples of model complex cells

In this section, we show examples of model complex cells trained on simple cell responses using natural images with temporal information. We demonstrate that even though efficient coding can pool simple cells with different spatial phase preferences that are sufficient to contribute to the spatial phase invariance, the model complex cells display no spatial phase invariance. The following examples account for the diversity observed in the population of model complex cells.

A complex cell that pools simple cells that have the same orientation but different phase preferences: Figure 3.20 shows the subspace of complex cell C86 where the simple cells in the subspace have the same orientation but different spatial phases. Unlike Figure 3.10, where simple cells only cover a limited region of spatial phase, simple cells in the subspace of complex cell C86 span over all possible spatial phases, suggesting that the subspace should generate spatial phase invariance. However, the spatial phase tuning curve of complex cell C86 does not show much spatial phase invariance with $F_1/F_0 = 1.1$. Nevertheless, the default linear model exhibits spatial phase invariance with a small value of F_1/F_0 (0.224).

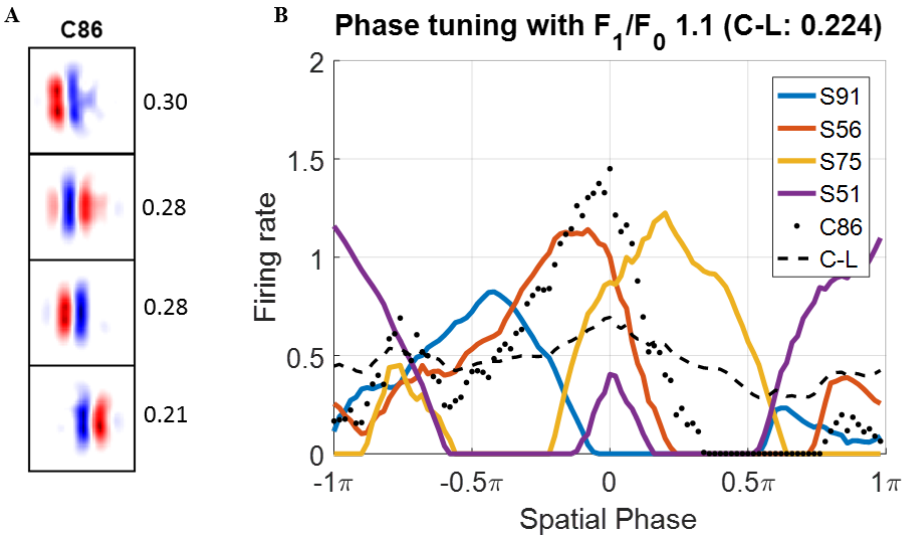

Figure 3.20: Complex cell C86. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C86. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the control model of complex cells that linearly sums simple cell responses weighted by the connection weights.

A complex cell that pools simple cells with two distinct orientations: Figure 3.21 shows an example of complex cell, C22, where the simple cells in the subspace have

two distinct orientations. As seen from Figure 3.21B, different simple cells are tuned to different spatial phases when sinusoidal gratings with preferred orientation and frequency are used as the input stimuli to the model. The subspace has two orientations, with the first (S64), third (S69), and fifth (S20) simple cells being more dominant as seen from their much stronger spatial phase tuning curves than the second (S75) and fourth (S99) simple cells. Though simple cells in the subspace do not cover all 2π radians) region of spatial phase, they contribute to a large extent of spatial phase invariance as can be seen from the spatial phase tuning curve of the default linear model that has $F_1/F_0 = 0.742$. However, the model complex cell exhibits very limited spatial phase invariance with $F_1/F_0 = 1.56$.

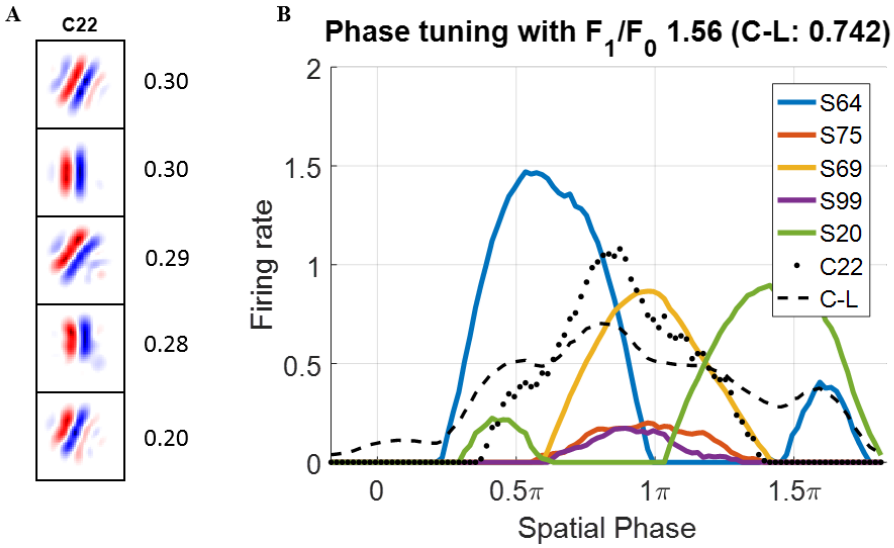

Figure 3.21: Complex cell C22. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C22. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the control model of complex cells that linearly sums up simple cell responses weighted by the connection weights.

A complex cell that pools simple cells with similar orientations: Figure 3.22 shows another example of complex cell, C5, that lies between the first and second examples: the simple cells in the subspace have different but similar orientations. As seen from Figure 3.22B, for the sinusoidal gratings with preferred orientation and frequency and different spatial phases, these simple cells in the subspace have different spatial phase tuning curves with different spatial phase preferences, which cover almost the whole 2π radians region of spatial phase. Therefore, the F_1/F_0 ratio is 0.416 for the default linear model, which indicates that the subspace is sufficient to generate spatial phase invariance. However, the spatial phase tuning curve of the model complex cell is limited to a subset

of 2π with $F_1/F_0 = 1.39$.

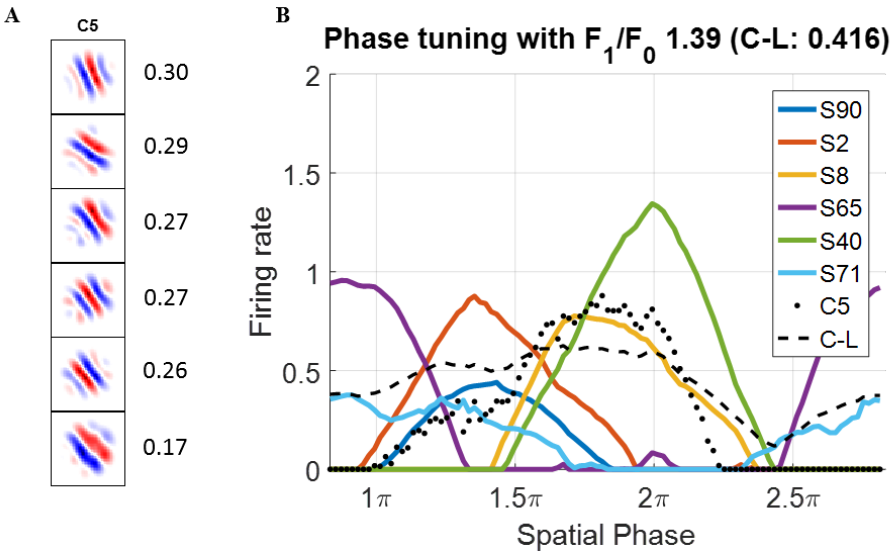

Figure 3.22: Complex cell C5. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C5. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dots represent the firing rates of the complex cell in response to different spatial phases. The dashed line is for the default linear model for the complex cell. S represents simple cell and the following number is the index of the simple cell. C-L represents the control model of complex cells that linearly sums up simple cell responses weighted by the connection weights.

Histogram of F_1/F_0 implies that efficient coding makes model complex cells 'simple'

The histogram of F_1/F_0 for the default linear model (Figure 3.23C) is closer to experimental data (Figure 3.23A) and shows a distribution centered at around 0.6 with most values smaller than 1, suggesting that the learned subspace of the model complex cell is actually sufficient to generate spatial phase invariance if the model is just a weighted linear summation of simple cell responses. However, the distribution for the model complex cells (Figure 3.23B) is skewed toward 2, indicating that most model complex cells are actually categorized as simple cells according to their values of F_1/F_0. Therefore, efficient coding makes model complex cells 'simple' by the principle of efficient coding itself because of suppressing model cell activities.

The above indicates that efficient coding on simple cell responses using natural images with temporal information can pool simple cells to form the subspace of complex cells. However, the competition between complex cells brought about by efficient coding suppresses complex cell responses such that they do not show spatial phase invariance.

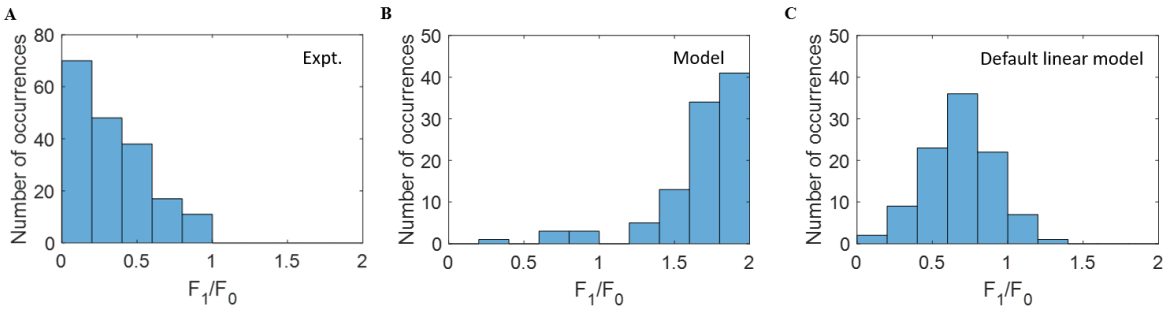

Figure 3.23: Histograms of F_1/F_0 with $\lambda_c = 0$ and 100 complex cells trained on natural images with jitter. (A) Experimental complex cells (Ringach et al., 2002). (B) Model complex cells. (C) default linear model of complex cells.

Problems of the current model

Apparently, the current model has the problem of suppressing model cell responses such that model complex cells become spatial phase selective. Possible reasons are discussed in the next section and a model that solves this problem is proposed in Chapter 4.

3.4 Discussion

The efficient coding model used in this chapter comes from the variant of sparse coding that incorporates some biological constraints. However, different models of implementing efficient coding should lead to similar results irrespective of the details of how it is implemented.

3.4.1 Static natural images vs. natural images with jitter

The results presented here suggest that static natural images are not sufficient to learn the subspace with simple cells that have diverse spatial phase tuning properties, while natural images with jitter help the model pool more simple cells with different spatial phase tuning preferences. As discussed earlier in the Methods Section, natural images with jitter have similar features to those with shift and other translations, which is similar to the temporal information contained in natural videos, so the temporal information is key for the model to selectively pool simple cells with similar orientations but various spatial phase preferences.

Efficient coding enables a model to find the efficient representation of the input, which also reduces the average activities of model units. Therefore, the learning rule of efficient coding helps model units capture the underlying features of the input, such as Gabor-like filters for natural images (Olshausen and Field, 1996, 1997; Van Hateren and Van Der Schaaf, 1998; Wiltschut and Hamker, 2009; Lian et al., 2019). For a simple cell, its response represents how strongly a feature, such as an orientated bar with preferred spatial

phase, is present in the input. Therefore, for a static natural image patch, responses of different simple cells represent the existence of different features, Gabor filters with different orientations, spatial frequencies, and spatial phases. However, in a natural image patch, Gabor filters with similar orientations but different spatial phases tend not to be co-active because oriented textures in one natural image patch only have one preferred spatial phase. Specifically, two simple cells with opposite spatial phase tuning properties (i.e., the difference between preferred spatial phases is π in radians) can never be co-active. Furthermore, corners and curves are prevalent in natural images, so some simple cells with RFs of different orientations will be simultaneously active. Because the principle of efficient coding pools co-active simple cells to efficiently represent simple cell responses, the learned subspace consists of simple cells with limited spatial phase invariance (Figure 3.11 and 3.10) different orientations (Figure 3.12), or a mixture of similar orientations (Figure 3.13).

In contrast, natural images with jitter contain more spatial phase information than static images because patches around a location tend to have similar patterns but shifted spatial phases. Therefore, the average activity of a simple cell will be comparable to other simple cells with similar orientations. As a result, efficient coding pools co-active simple cells into the subspace, which consequently has more diversity in spatial phase tuning.

3.4.2 The trade-off between selectivity and competition

The principle of efficient coding and its associated learning rule helps the model units learn features, which also makes model units highly selective to their preferred features. The competition brought about by efficient coding is indispensable for achieving selectivity and diversity of response, either through feedback (Wiltschut and Hamker, 2009; Lian et al., 2019) or lateral connections (Rozell et al., 2008; Zylberberg et al., 2011). However, competition might also be very strong, as can be seen from Figure 11 of Olshausen and Field (1997), where the feedforward response is much stronger than the response in the efficient coding model. Different model complex cells may pool the same simple cell, but complex cells might lose the spatial phase invariance brought by the simple cell when complex cells are competing with each other to represent the simple cell response. Therefore, the spatial phase tuning curves of such model complex cells are much narrower than the default linear model, which implies that efficient coding help model complex cells selectively pool simple cells, but that the competition introduced by efficient coding suppress model cell responses such that they behave like simple cells.

Why does zero sparsity level still introduce strong competition?

The model trained on natural images with jitter has the zero sparsity level for complex cells ($\lambda_C = 0$), but the model cell responses are highly suppressed compared with the

default linear model (Figure 3.23). Why does λ_C still introduce competition into the network? There are two reasons: (1) though $\lambda_C = 0$, the firing mechanism is still a one-side nonlinearity that rectifies negative values, which means model cells need to compete to be positive in order to represent the input; (2) The efficient coding model implemented here has the same form as predictive coding that was shown to have a similar structure to the competition model (Spratling, 2008), so the model naturally introduces competition.

3.4.3 Why other efficient coding models of complex cells can achieve spatial phase invariance?

Hyvärinen and Hoyer (2000 and 2001) used efficient coding to learn a subspace that can achieve spatial phase invariance of complex cells. However, these models require some biologically unrealistic assumptions. First, the quadratic pooling function (double-sided nonlinearity) assigns the subspace with polarity invariance (π spatial phase invariance). Therefore, their models learn spatial phase invariant subspace using static natural images. Second, the dynamics of their model cells is a local computation that linearly sums responses of subunits, but the learning process uses global information from other model cells.

Therefore, when we add these biological constraints and apply efficient coding on complex cells, we failed to achieve spatial phase invariance.

3.5 Conclusion

In this chapter, a model of complex cells was investigated to see if it could explain spatial phase invariance of complex cells using the principle of efficient coding, where simple cell responses were used as the input to the complex cells. We found that the model could not pool simple cells with different spatial phase tuning to form the subspace of a complex cell trained on static natural images. When natural images with jitter that contain features of temporal information were used to train the model, the learned subspace of complex cells consisted of simple cells with different spatial phase tuning properties that were covered a wide range of spatial phases, as would be expected for phase invariant complex cells. However, the principle of efficient coding suppressed complex cell responses to many spatial phases, with the result that complex cell responses did not show the expected range of spatial phase invariance.

Chapter 4

Learning receptive field properties of complex cells

4.1 Introduction

In this chapter, learning by complex cells in a biologically plausible model within the hierarchical structure is further investigated. Pooled simple cells form the *subspace* of the complex cell and each pooled simple cell is a *subunit* in the subspace.

The results of Chapter 3 indicated that temporal information is important for pooling the simple cells into the subspace of a complex cell in primary visual cortex (V1), but that efficient coding failed to produce complex cell properties because the strong competition suppressed the response to some units in the subspace of complex cells.

In contrast, another learning rule - the Bienenstock, Cooper and Munro (BCM) plasticity rule (Bienenstock et al., 1982; Cooper and Bear, 2012) - can also learn underlying features from the input through a competitive process that arises from the thresholding mechanism that is part of the BCM learning rule, but the BCM plasticity is designed for single unit and does not introduce any competition between network units that receive the same visual input. Law and Cooper (1994) applied the BCM plasticity rule to a network using natural images as input stimuli and showed that this learning rule can learn simple cell-like receptive fields (RFs). However, since the BCM plasticity rule is the same for every neuron in the network, the learned features of the network tend to be similar (Willmore et al., 2012). By incorporating contrast normalization, namely where the response of a cell is normalized by responses of other cells in the network Heeger (1992), a "soft" form of competition is introduced to the network. Furthermore, increasing experimental evidence has been found to support the hypothesis that normalization operates throughout the visual system, from the retina to the visual cortex (see Carandini and Heeger (2012) for a review). By incorporating BCM plasticity and normalization, Willmore et al. (2012) showed that normalized BCM (NBCM) can learn different simple cell-like RFs when the

model is trained on natural images. However, the BCM and NBCM plasticity rules ignore some biological constraints such as positive neuronal response and Dale's Law.

In this chapter, a biologically plausible learning model for complex cells to pool simple cells using the modified version of the BCM plasticity rule is proposed demonstrating that, when the model is trained on natural images, model complex cells can pool simple cells with various spatial phase preferences into a subspace that can account for the spatial phase invariance of experimental complex cells. However, the subspaces of different model complex cells are highly repetitive. To overcome this problem, it is shown that a modified version of the BCM rule, namely the normalized-BCM (NBCM) rule, can learn different complex cells and so respond to different features of visual input. Further analysis on model complex cells shows that the proposed model can account for the diversity of receptive field (RF) properties of complex cells found in a recent experimental study (Almasi, 2017).

4.2 Methods

4.2.1 Structure of the model

The proposed three-layer network of rate-based neurons models the activities of lateral geniculate nucleus (LGN) cells (first layer), V1 simple cells (middle layer) and V1 complex cells (top layer), respectively, as shown in Figure 4.1. A summary of the parameters of the model that will be used throughout this chapter is given in Table 4.1.

The bottom two layers implement the two-layer model that is biologically plausible and can account for many experimental phenomena (Chapter 2, Lian et al. (2019)). The dynamics of LGN cells and simple cells are given by (the same as Eq. 3.1 and 3.2)

$$\begin{aligned} \tau_{\mathrm{L}}\dot{\mathbf{v}}^{\mathrm{L}} &= -\mathbf{v}^{\mathrm{L}} + \mathbf{x}^{\mathrm{L}} + (\mathbf{A}^{\mathrm{d},+} + \mathbf{A}^{\mathrm{d},-})\mathbf{r}^{\mathrm{S}} + r_{\mathrm{b,L}} \\ \mathbf{r}^{\mathrm{L}} &= \max(\mathbf{v}^{\mathrm{L}}, 0), \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} \tau_{\mathrm{S}}\dot{\mathbf{v}}^{\mathrm{S}} =& -(\mathbf{v}^{\mathrm{S}} - \mathbf{v}^{\mathrm{S}}_{\mathrm{leak}}) + \mathbf{A}^{\mathrm{u},+T}_{\mathrm{ON}}\mathbf{r}^{\mathrm{L}}_{\mathrm{ON}} + \mathbf{A}^{\mathrm{u},-T}_{\mathrm{ON}}\mathbf{r}^{\mathrm{L}}_{\mathrm{ON}} \\ &+ \mathbf{A}^{\mathrm{u},+\ T}_{\mathrm{OFF}}\mathbf{r}^{\mathrm{L}}_{\mathrm{OFF}} + \mathbf{A}^{\mathrm{u},-\ T}_{\mathrm{OFF}}\mathbf{r}^{\mathrm{L}}_{\mathrm{OFF}} + \mathbf{r}^{\mathrm{S}} \\ \mathbf{r}^{\mathrm{S}} =& \max(\mathbf{v}^{\mathrm{S}} - \lambda_{\mathrm{S}}, 0). \end{aligned} \tag{4.2}$$

Details of the model of the bottom two layers can be found in Chapter 3 (Section 3.2).

The dynamics of the complex cells is simply the linear summation of simple cells they connect, as given by

$$\mathbf{r}^{\mathrm{C}} = \mathbf{A}^{\mathrm{C}T}_{\mathrm{S}}\mathbf{x}^{\mathrm{C}} \tag{4.3}$$

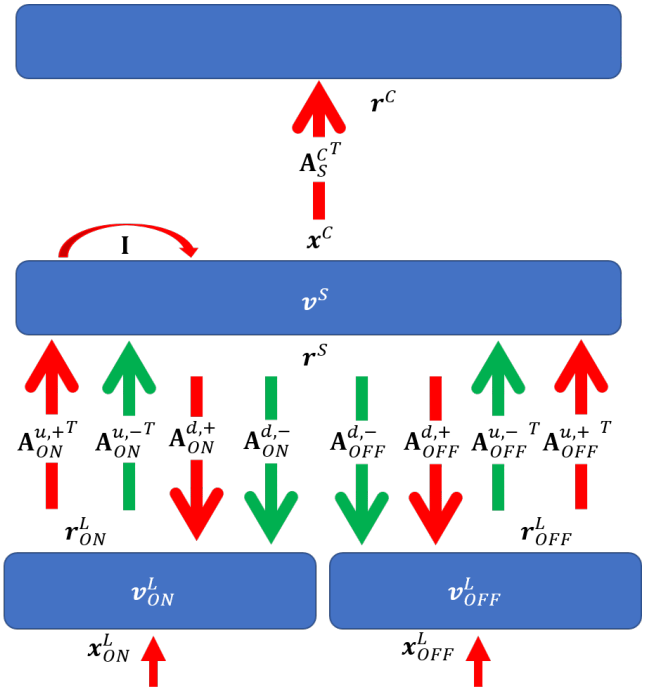

Figure 4.1: Graphical representation of the model. **I** is the identity matrix that represents self-excitation. Red and green arrows represent excitatory and inhibitory connections, respectively. Upward and downward arrows are for feedforward and feedback pathways. Parameters defined in Table 4.1.

where $\mathbf{A}_{\mathrm{S}}^{\mathrm{C}}$ is a non-negative matrix that represents excitatory connections between simple and complex cells and $\mathbf{x}^{\mathrm{C}}$ is the input to the complex cell.

4.2.2 Input and pre-processing procedure

It was shown in Chapter 3 that the input of natural images with jitter can integrate temporal information that helps complex cells pool simple cells with various spatial phase preferences. Therefore, the input to the model used in this chapter is natural images with jitter, where N patches around an image region are used as the input to simple cells and the average activity of simple cells in response to these N image patches is the input to complex cells (see details in Section 3.2). $\mathbf{x}^{\mathrm{C}}$ is the average activity of simple cells that incorporates the temporal information in the input stimuli; i.e., $\mathbf{x}^{\mathrm{C}} = \langle \mathbf{r}^{\mathrm{S}} \rangle$, where $\langle \cdot \rangle$ denotes the ensemble average, i.e., the average over multiple trials. The pre-processing procedure in this chapter is the same as Chapter 3 (see details in Section 3.2).

4.2.3 Learning rule for LGN-simple cell connection

Connections between the LGN and simple cells are learned based on Hebbian or anti-Hebbian plasticity, as described in Chapter 3 (Section 3.2). The learning rules (the same

Table 4.1: Model symbols and parameters in Chapter 4.

Description	**Symbol**
Input stimuli to LGN/complex cells	$\mathbf{x}^{\mathrm{L}}$ / $\mathbf{x}^{\mathrm{C}}$
Input stimuli to ON/OFF LGN cells	$\mathbf{x}^{\mathrm{L}}_{\mathrm{ON}}$ / $\mathbf{x}^{\mathrm{L}}_{\mathrm{OFF}}$
Membrane time constant of LGN/simple cells (10 ms)	τ_{L} / τ_{S}
Membrane potentials of LGN/simple cells	$\mathbf{v}^{\mathrm{L}}$ / $\mathbf{v}^{\mathrm{S}}$
Membrane potentials of ON/OFF LGN cells	$\mathbf{v}^{\mathrm{L}}_{\mathrm{ON}}$ / $\mathbf{v}^{\mathrm{L}}_{\mathrm{OFF}}$
Firing rates of LGN/simple/complex cells	$\mathbf{r}^{\mathrm{L}}$ / $\mathbf{r}^{\mathrm{S}}$ / $\mathbf{r}^{\mathrm{C}}$
Firing rates of ON/OFF LGN cells	$\mathbf{r}^{\mathrm{L}}_{\mathrm{ON}}$ / $\mathbf{r}^{\mathrm{L}}_{\mathrm{OFF}}$
Spontaneous firing rate of LGN (0.5 Hz)	$r_{\mathrm{b,L}}$
Leakage voltages of simple cells	$\mathbf{v}^{\mathrm{S}}_{\mathrm{leak}}$
Excitatory connection: all LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},+}$
Excitatory connection: ON/OFF LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},+}_{\mathrm{ON}}$ / $\mathbf{A}^{\mathrm{u},+}_{\mathrm{OFF}}$
Inhibitory connection: all LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},-}$
Inhibitory connection: ON/OFF LGN cells to simple cells	$\mathbf{A}^{\mathrm{u},-}_{\mathrm{ON}}$ / $\mathbf{A}^{\mathrm{u},-}_{\mathrm{OFF}}$
Excitatory connection: simple cells to all LGN cells	$\mathbf{A}^{\mathrm{d},+}$
Excitatory connection: simple cells to ON/OFF LGN cells	$\mathbf{A}^{\mathrm{d},+}_{\mathrm{ON}}$ / $\mathbf{A}^{\mathrm{d},+}_{\mathrm{OFF}}$
Inhibitory connection: simple cells to all LGN cells	$\mathbf{A}^{\mathrm{d},-}$
Inhibitory connection: simple cells to ON/OFF LGN cells	$\mathbf{A}^{\mathrm{d},-}_{\mathrm{ON}}$ / $\mathbf{A}^{\mathrm{d},-}_{\mathrm{OFF}}$
Excitatory connection: simple cells to complex cells	$\mathbf{A}^{\mathrm{C}}_{\mathrm{S}}$
Sparsity level (0.1)	λ_{S}
Upper bounds of LGN-simple/simple-complex connection weights (0.3 and 5)	$a_{1,\max}$ / $a_{2,\max}$
Learning rate of LGN-simple connection weights (3)	η_1
Learning rate of weights and thresholds (resp.) for complex cells (0.001 and 0.01)	η_a / η_θ
Weight regulation constants of LGN-simple/simple-complex connections (both 10^{-4})	γ_1 / γ_a
Parameters of normalization of complex cells (0.01 and 12)	α / β

as Eq. 3.6) are given by

$$\begin{aligned}
\Delta\mathbf{A}^{\mathrm{u},+} &= \eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{u},+} \right) \\
\Delta\mathbf{A}^{\mathrm{u},-} &= \eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{u},-} \right) \\
\Delta\mathbf{A}^{\mathrm{d},+} &= -\eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{d},+} \right) \\
\Delta\mathbf{A}^{\mathrm{d},-} &= -\eta_1 \left(\langle (\mathbf{r}^{\mathrm{L}} - r_{\mathrm{b,L}}) \mathbf{r}^{\mathrm{S}T} \rangle - \gamma_1 \mathbf{A}^{\mathrm{d},-} \right);
\end{aligned} \tag{4.4}$$

the details can be found in Section 3.2.

4.2.4 Learning rule for simple-complex cell connections

Connections between simple and complex cells are learned based on the BCM rule, a form of Hebbian plasticity where the efficacy change depends not only on pre- and post-synaptic activities but also on the slow varying values of the history of post-synaptic activities (Bienenstock et al., 1982).

For the synaptic weight between simple cell i and complex cell j, $a_{i,j}$, BCM learning rule updates the weight according to not only pre- and post-synaptic activities, x_i^{C} and r_j^{C}, but also a learned threshold for complex cell j, θ_j:

$$\begin{aligned} \Delta a_{i,j} &= \eta_a x_i^{\mathrm{C}} r_j^{\mathrm{C}} (r_j^{\mathrm{C}} - \theta_j) \\ \Delta \theta_j &= \eta_\theta \left((r_j^{\mathrm{C}})^2 - \theta_j \right), \end{aligned} \tag{4.5}$$

where η_a and η_θ are the learning rates that determine the rates of change for the synaptic weight and threshold. Original BCM rule allows weight $a_{i,j}$ to change sign.

Modified BCM rule

For the modified BCM rule, the synaptic weight between simple cell i and complex cell j, $a_{i,j}$, is updated by the learning rule:

$$\begin{aligned} \Delta a_{i,j} &= \eta_a \left(x_i^{\mathrm{C}} r_j^{\mathrm{C}} (r_j^{\mathrm{C}} - \theta_j) - \gamma_a a_{i,j} \right) \\ \Delta \theta_j &= \eta_\theta \left((r_j^{\mathrm{C}})^2 - \theta_j \right), \end{aligned} \tag{4.6}$$

where γ_a is the weight regulation constant. In addition, the connections between simple and complex cells are excitatory in the model, so $a_{i,j}$ is kept non-negative during learning. The upper bound of the connection weights is explicitly constrained to be $a_{2,\max}$.

Note that the **modified BCM** (Eq. 4.6) learning rule differs from the original BCM (Eq. 4.5) learning rule in three ways. First, the original BCM rule allows weights to change signs, which is not permitted in the modified BCM rule. Second, $a_{i,j}$ is constrained by the maximal weight, $a_{2,\max}$. Third, the original BCM rule does not have the weight regulation term, $-\gamma_a a_{i,j}$. This term is added to prevent weights from growing without bound and to push unimportant weights to zero.

Modified NBCM rule

The **original NBCM** rule proposed by Willmore and colleagues shows that this learning rule can learn different RFs for neurons in a network (Willmore et al., 2012). NBCM incorporates the response normalization model proposed by Heeger (1992). For the model,

the response of complex cells, r_j^{C}, is normalized and the normalized response, $r_{j,N}^{\mathrm{C}}$, is then used to update the synaptic weight and threshold, as given by NBCM learning rule is given by

$$\begin{aligned} r_{j,N}^{\mathrm{C}} &= \frac{\beta r_j^{\mathrm{C}}}{\alpha + \sqrt{\sum_k (r_k^{\mathrm{C}})^2}} \\ \Delta a_{i,j} &= \eta_a x_i^{\mathrm{C}} r_{j,N}^{\mathrm{C}} (r_{j,N}^{\mathrm{C}} - \theta_j) \\ \Delta \theta_j &= \eta_\theta \left((r_{j,N}^{\mathrm{C}})^2 - \theta_j \right), \end{aligned} \tag{4.7}$$

where k represents any possible index of complex cells in the network and α and β are constants that determine the strength of the normalized response, $r_{j,N}^{\mathrm{C}}$, compared with response, r_j^{C}.

The **modified NBCM** learning rule introduces more constraints on the weights and is given by

$$\begin{aligned} r_{j,N}^{\mathrm{C}} &= \frac{\beta r_j^{\mathrm{C}}}{\alpha + \sqrt{\sum_k (r_k^{\mathrm{C}})^2}} \\ \Delta a_{i,j} &= \eta_a \left(x_i^{\mathrm{C}} r_{j,N}^{\mathrm{C}} (r_{j,N}^{\mathrm{C}} - \theta_j) - \gamma_a a_{i,j} \right) \\ \Delta \theta_j &= \eta_\theta \left((r_{j,N}^{\mathrm{C}})^2 - \theta_j \right), \end{aligned} \tag{4.8}$$

where γ_a is the weight regulation constant. Additionally, maximal value of the connection weights is explicitly constrained to be $a_{2,\max}$. As above, the **modified NBCM** learning rule (Eq. 4.8) differs from the original NBCM learning rule (Eq. 4.7) in three ways: $a_{i,j}$ is kept non-negative during learning, $a_{i,j}$ is constrained by the maximal connection weight, $a_{2,\max}$, and there is a weight regularization term.

It should be mentioned that the normalization equation in Eq. 4.8 is not obviously physiologically plausible because it uses the global information of other neuron activities to calculate normalized responses of the post-synaptic neuron. However, such normalization is consistent with a lot of experimental data (Carandini and Heeger, 2012). Furthermore, how this normalization can arise in a biologically plausible fashion has been shown in the supralinear stabilized network model through lateral connections with physiologically realistic neural dynamics involving recurrent connections within V1 (Rubin et al., 2015).

4.2.5 Training

Natural images of size 16×16 pixel are used as the input. The training process for simple cells (bottom two layers) is the same as Chapter 3 (Section 3.2). After the first two layers are learned, connection weights between LGN and simple cells are fixed and the process of learning complex cells starts. The models described in this chapter are implemented in MATLAB (version R2016b, MathWorks, MA, USA) using my own codes.

Applying modified BCM rule for complex cells

For the modified BCM rule, there are 10^6 epochs in the training process. The learning rates for weights and threshold are $\eta_a = 0.001$ and $\eta_\theta = 0.01$, respectively, similar to values in Willmore et al. (2012). The weight regularization constant, γ_a, is set to 0.0001 and the maximal connection weight, $a_{2,\max}$, is 5. The number of image patches, N, is taken to be 20.

Applying the modified NBCM rule for complex cells

For the modified NBCM rule, there are 10^6 epochs in the training process. The learning rates for weights and threshold are 0.001 (η_a) and 0.01 (η_θ), respectively. The parameters for the divisive normalization in Eq. 4.8 are $\alpha = 0.01$ and $\beta = 12$. Similarly, the weight regularization constant, γ_a, is chosen to be 0.0001 and the maximal connection weight, $a_{2,\max}$, is set to 5. The number of image patches, N, is taken to be three different values: 5, 10 and 20.

4.2.6 Determining the level of repetition among model complex cells

After learning, each model complex cell selectively pools some simple cells with different weights. In order to investigate whether the model learns different complex cells, several measurements are used.

First, simple cells that have pronounced weights (> 2, the maximal weight allowed is 5) are investigated. The mean, standard deviation, and maximum of the number of complex cells that have pronounced connection weights with each simple cell are calculated. The number of simple cells that have no pronounced connection with any complex cells are also recorded.

Then, a k-mean clustering algorithm with $k = 25$ is used to cluster complex cells into 25 groups based on the connections weights of each complex cell. Similar complex cells should be clustered into the same group by this algorithm. If there are many complex cells in the same group, it indicates that the learned complex cells are highly repetitive.

4.2.7 Measuring spatial phase invariance

Similar to Chapter 3 (Section 3.2.5), the spatial phase invariance for model complex cells is investigated using the spatial phase tuning curve and F_1/F_0 ratio when sinusoidal gratings with the preferred orientation and frequency but different phases are presented as the input stimuli to the model. Details can be found in Section 3.2 and Appendix A.

4.2.8 Measuring orientation tuning

Sinusoidal gratings, with preferred frequency and all possible phases spanning 0 to 360 degrees with step size of 3.6 degrees, were presented to the model. The maximal response, r_k, to the gratings with all possible phases and orientation ϕ_k is recorded. The orientations ϕ_k range from 0 to 360 degrees with a step size of 3.6 degrees. Therefore, there are 100 pairs of r_k and ϕ_k for each model complex cell, which generates the orientation tuning curve.

Circular variance: The first measure of orientation tuning is *circular variance* that provides a robust bounded index ranging from 0 to 1 (Mardia, 1972; Batschelet, 1981; Ringach et al., 2002). The circular variance, C_v, is calculated as

$$C_v = 1 - |R|, \text{ where } R = \frac{\sum_k r_k \ e^{i2\phi_k}}{\sum_k r_k}, \tag{4.9}$$

where k represents any possible discrete index of orientations. Note that the unit of orientation, ϕ_k, should be converted to radians when computing C_v in Eq. 4.9. If a cell has no orientation selectivity, namely that the cell responds equally to different orientation, the circular variance is 1. However, the other extreme is that this cell only responds to a specific orientation, where the circular variance is 0. Therefore, highly orientation-selective cells are mapped to C_v close to 0 and weakly orientation-selective cells are mapped to C_v close to 1.

Half-bandwidth: The second measure of orientation tuning is the *half-bandwidth* at $1/\sqrt{2}$ height of the orientation tuning curve. This measure is similar to the standard deviation of a Gaussian distribution, indicating how widely the cell is tuned to the preferred orientation. Half-bandwidth was also used in previous experimental studies (Schiller et al., 1976b; Ringach et al., 2002).

4.2.9 Analyzing complex cells using nonlinear input model

Nonlinear input model

The nonlinear input model (NIM) proposed by McFarland et al. (2013) was used to reveal how underlying inputs might contribute to a neuron's response and to compare with experimental data. Almasi and colleagues applied NIM to experimental data of cat V1 (Almasi, 2017) to analyze RF properties of complex cells and found that there is a wide range of complex cell response types in cat V1. Therefore, NIM is also used here to analyze RF properties of model complex cells so that a comparison between model and experimental data can be conducted.

Figure 4.2 shows the structure of NIM. NIM assumes a hierarchical structure, where the response is a nonlinear function of the sum of responses of different filters. The

response of the filter is determined by the corresponding linear filter and nonlinearity. The filters and upstream nonlinearities are fit non-parametrically, and the spiking nonlinearity is a monotonically increasing threshold function with the form

$$y(x) = \frac{\alpha}{\beta} \log(1 + e^{\beta(x-\theta)}), \tag{4.10}$$

where α, β, and θ are the parameters to be estimated. It should be mentioned that filters of NIM are not necessarily the simple cells (subunits) pooled by a complex cell, because simple cells are combined to form filters and nonlinearities of NIM. The fitting routine is done using maximum likelihood of the model parameters given the stimulus and response. Overall, NIM fitting only assumes a hierarchical structure and a form of spiking nonlinearity, while other components are determined by the stimulus (input) and response (output).

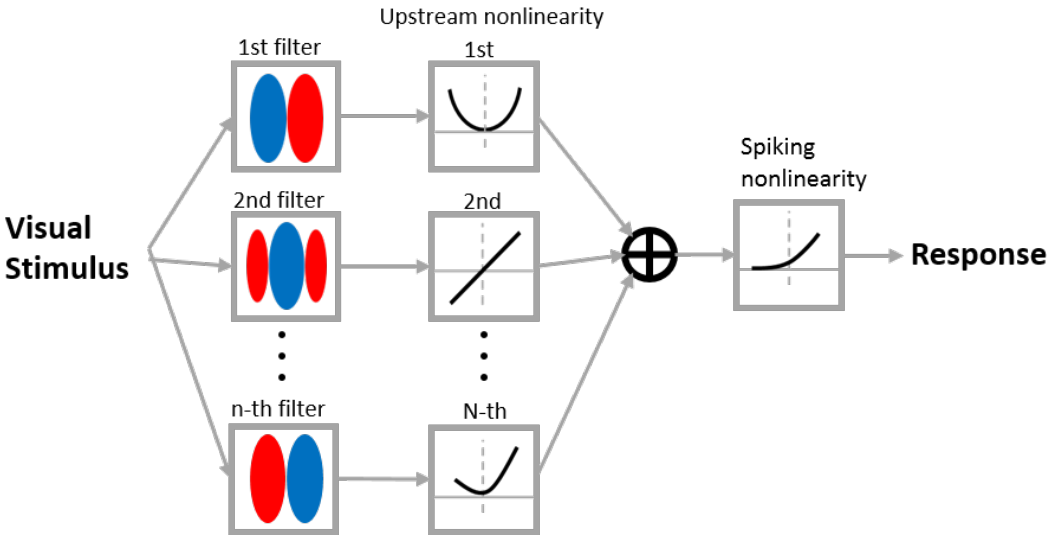

Figure 4.2: The structure of nonlinear input model (NIM). The filter and upstream nonlinearity determine how it responds to the visual input. The sum of responses of all filters are then passed to a spiking nonlinear function to generate the response of the model.

In the experimental study of Almasi (2017), responses of V1 cells were recorded while the anaesthetized cat was presented with images of white noise (pixel intensities randomly chosen according to a Gaussian distribution) and NIM was then applied onto pairs of images and responses to analyze RF properties. Therefore, in order to fit complex cells of the trained model using NIM in a similar fashion, 32×32 pixel images of white noise are generated, pre-processed, and input to the model and then responses of model complex cells are collected. Next, pairs of images and model responses are used in NIM fitting. It is assumed that the 32×32 pixel image is equivalent to 10 degrees of the visual field. In this way, the unit of spatial frequency of the model can be translated from cycles/pixel into

experimental unit, cycles/degree. The fitting routine[1] fits filters, upstream nonlinearities, and spiking nonlinearity simultaneously. For the purpose of comparing model results to complex cells that have two filters in the experimental study Almasi (2017), the number of filters is set to two when using the fitting routine for model complex cells.

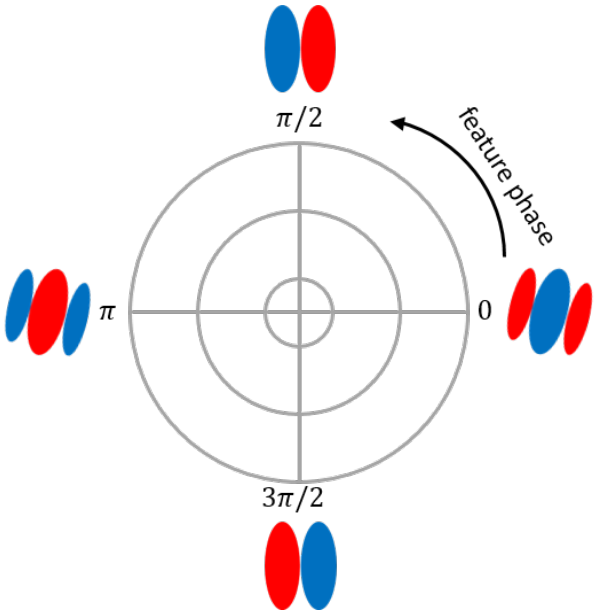

Figure 4.3: Illustration of an example 2-D feature spectrum spanned by the two filters fitted by NIM.

Feature phase bandwidth, orientation breadth, spatial frequency breadth, and spatial phase breadth

After NIM fitting, each complex cell has two filters (features) and two corresponding upstream nonlinearities. Then a 2-D feature spectrum (shown in Figure 4.3) is used to represent the space spanned by the two filters. By varying the feature phase, any point in the feature spectrum represents a feature composed by a linear combination of the two filters.

In order to investigate the tuning properties of spatial phase, orientation, and spatial frequency for a population of complex cells, *feature phase bandwidth*, *orientation breadth*, *spatial frequency breadth*, and *spatial phase breadth* are four quantitative measurements[2] used to analyze model data.

Feature phase bandwidth is the extent of the feature phase domain where the cell has responses larger than a threshold, and its value can vary between 0 to 2π. As feature

[1]The fitting routine was kindly provided by Almasi, A., Ibbotson, M., and Meffin, H. at National Vision Research Institute in Melbourne, Australia.

[2]These methods were developed by Almasi, A., Ibbotson, M., and Meffin, H. at National Vision Research Institute in Melbourne, Australia.

phase varies from 0 to 2π, the NIM model predicts the response for each feature phase. Feature phase bandwidth is just the sum of all the phase intervals where the responses are above the threshold. The threshold, r_θ, is taken to be $0.5(r^{\max} - r^{\text{base}}) + r^{\text{base}}$, where $r^{\max}$ and r^{base} are maximum and background responses respectively. Feature phase bandwidth is determined by the upstream nonlinearities. If the nonlinearities of the two filters are both quadratic functions, feature phase bandwidth will be close to 2π radians (360 degrees) because the response is always large as feature phase varies from 0 to 2π.

Orientation breadth, spatial frequency breadth and spatial phase breadth measure the range of variation in these three characteristics determined by the NIM model. For each feature phase, there is one unique interpolated feature with characteristics of orientation, spatial frequency and spatial phase. As feature phase varies from 0 to 2π, distributions of orientation, spatial frequency and spatial phase can obtained. However, only feature phases that invoke response larger than the threshold, r_θ, are used to generate the distributions of orientation, spatial frequency and spatial phase of the NIM model. Then orientation breadth, spatial frequency breadth and spatial phase breadth are just the extent of the domains that corresponding distributions cover. The maximal values of orientation breadth, spatial frequency breadth and spatial phase breadth were considered to be 180°, 0.8*cpd* (circles per degree) and 360°, respectively. In practice, the characteristics are discretised to get empirical values of the above measurements.

Large values of orientation breadth indicate that the cell shows invariance to perturbations in orientation. If a cell has two filters with distinct orientations, orientation breadth will be relatively large. Large values of spatial frequency breadth indicate the cell is invariant to a wider range of spatial frequencies. Since the values of spatial frequency depend on the size (in degrees) of the visual field that correspond to the input image, spatial frequency breadth only provides a qualitative rather than quantitative comparison between model and experimental data. Large values of spatial phase breadth indicate strong spatial phase invariance. For the classical energy model, spatial phase breadth will be close to 2π in radians (360 degrees).

4.3 Results

4.3.1 The model based on the modified BCM rule

The model can learn the subspaces of complex cells

Two examples of model complex cells, C3 and C96, are displayed in Figures 4.4 and 4.5 to illustrate that the model can learn spatial phase invariance while still keeping orientation selectivity. In order to show which simple cells provide pronounced input to each complex cell, only simple cells with connection weights larger than 2 (the maximal

value of the weight is 5) are displayed. These will be referred to as the pronounced simple cell inputs. Note that the firing rates of complex cells are divided by $a_{2,\max}$ in order to show the responses of simple and complex cells in the similar range (Figure 4.4B,C and Figure 4.5B,C). Figures 4.4 and 4.5 show that, for each complex cell, the pronounced simple cell inputs have similar orientations but widely different spatial phases, which makes the model cells invariant to spatial phase but still selective to orientation.

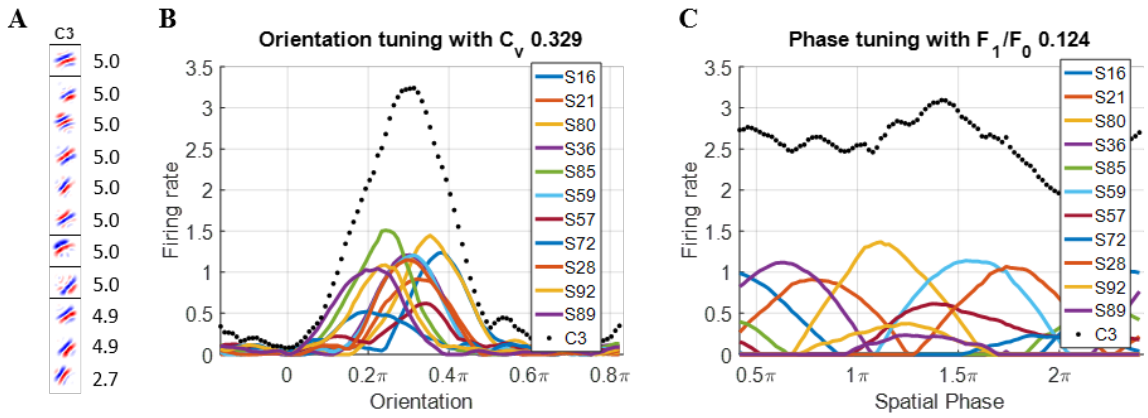

Figure 4.4: Complex cell C3. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C3. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Orientation tuning curves. (C) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dotted line is for complex cell C3. S represents simple cell and the following number is the index of the simple cell.

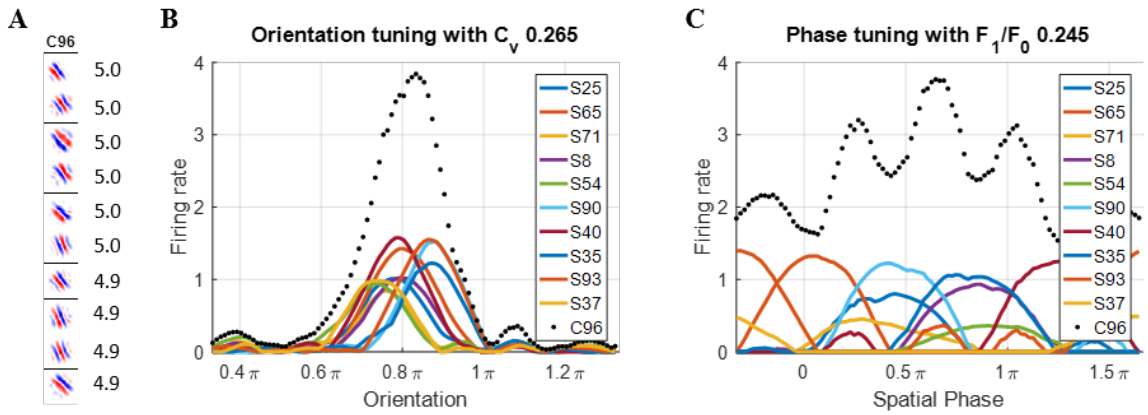

Figure 4.5: Complex cell C96. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C96. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Orientation tuning curves. (C) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dotted line is for complex cell C96. S represents simple cell and the following number is the index of the simple cell.

Many model complex cells are similar

Figures 4.4 and 4.5 show results from two representative model cells that are invariant to spatial phase but selective to orientation. However, the model complex cells are not diverse but instead have similar subspaces (full subspaces are displayed in Figure B.4 of Appendix B). The scatter plot of simple-complex cell connections is shown in Figure 4.6 where each dot indicates that the connection weight between the simple and complex cell has a pronounced weight (> 2). As seen in Figure 4.6, many complex cells have pronounced connections with the same simple cells, so dots in the scatter plot form some vertical lines. However, some simple cells have no pronounced connection with any complex cell.

For 100 simple cells in the modified BCM rule model, each simple cell is pooled by on average 10.2 complex cells with standard deviation 18.6. Furthermore, 56 simple cells are not connected to any complex cells, while the maximum number of complex cells connected with the same simple cell is 76. Therefore, complex cells learned by the BCM plasticity rule are highly repetitive by having the same simple cells in the subspace. In addition, after k-means clustering with $k = 25$ is performed to cluster 100 model complex cells into 25 different groups based on the connections weights with simple cells, one cluster has 48 complex cells, indicating that these complex cells are very similar.

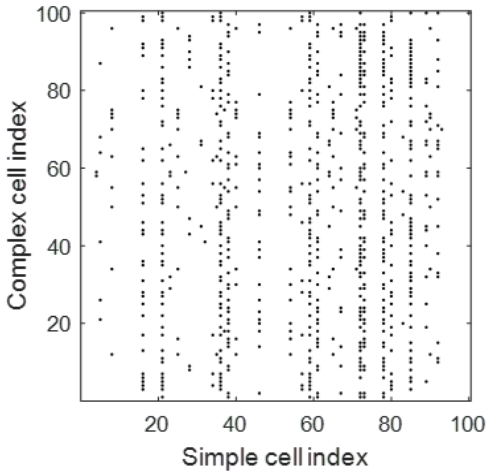

Figure 4.6: Scatter plot of simple-complex cell connections for the model based on modified BCM rule. The dots in each row represent the indices of simple cells that have pronounced weights (> 2) with the complex cell.

4.3.2 The model based on the modified NBCM learning rule

Different subspaces of complex cells are learned

In comparison with model complex cells learned by the modified BCM rule, the modified NBCM rule helps the model to learn different complex cells with different orientations

such that there is little overlap and repetition in pronounced simple cell inputs across the complex cell population (full subspaces can be found in Figure B.5 of Appendix B). The scatter plot of simple-complex cell connections is shown in Figure 4.7, which shows that simple-complex connections are more diverse than given by the modified BCM learning rule (Figure 4.6). Dots in the figure are more random and there are no vertical lines that appear in Figure 4.6 that indicate many connections from individual simple cells to many complex cells.

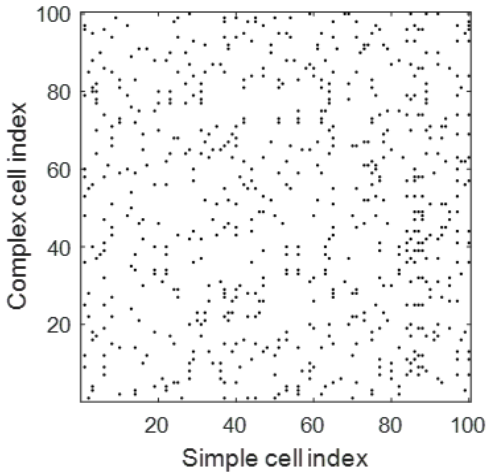

Figure 4.7: Scatter plot of simple-complex cell connections for the model based on learning with the modified NBCM rule. The dots in each row represent the indices of simple cells that have pronounced weights (> 2) with the complex cell.

For 100 simple cells in the modified NBCM rule model, each simple cell is pooled by on average 8.8 complex cells with standard deviation 4.8. The standard deviation is much smaller than that of the modified BCM rule. Furthermore, only 1 simple cell is not connected with any complex cells compared with 56 unconnected simple cells for the modified BCM rule. The maximum number of complex cells connected to the same simple cell is only 20, which is much smaller than 76 for the modified BCM rule. In addition, after k-means clustering with $k = 25$ is performed to cluster 100 model complex cells into 25 different groups based on the connections weights with simple cells, the largest cluster has 8 complex cells. Above all, the model complex cells learned by the modified NBCM rule tend to be more diverse.

Number of image patches determines the level of spatial phase invariance of model complex cells

In this section, I first show some population statistics of F_1/F_0, circular variance, and half-bandwidth using three different values for the number of image patches, defined as N, ($N = 5,\ 10,$ and 20) for the modified NBCM rule model with parameters given in

Table 4.1. The model responses and are derived from simulations of drifting gratings (see Section 4.2.7 and 3.2.5 and for details).

Intuitively, the larger N is, the more spatial phase invariant the model complex cells will be because more image patches are taken into consideration and so more spatial phases of the same features will be sampled. This is consistent with the simulation results, as can be seen in Figure 4.8, which shows that the distribution of F_1/F_0 skews towards zero when N increases, indicating greater spatial phase invariance.

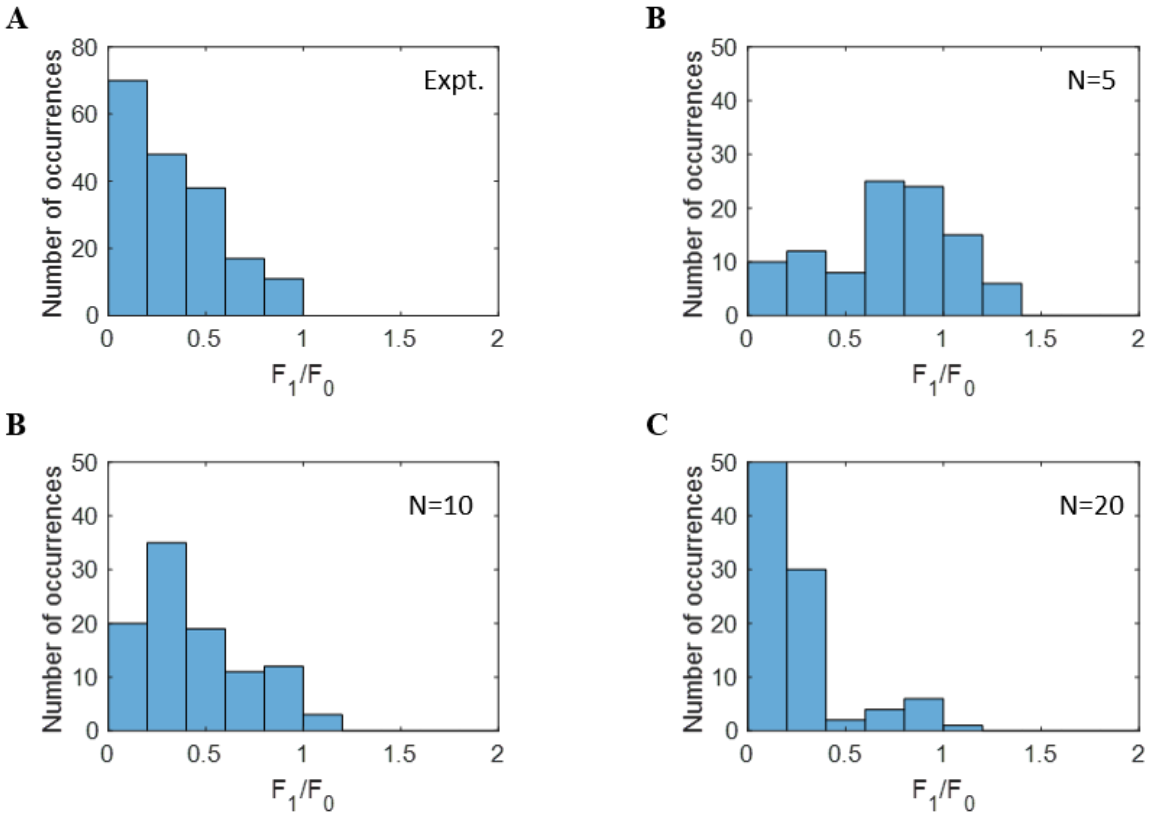

Figure 4.8: Histograms of F_1/F_0. (A) Experimental complex cells. Model complex cells with (B) $N = 5$, (C) $N = 10$, and (D) $N = 20$.

Figure 4.9 shows scatter plots of F_1/F_0 vs. C_v (defined in Methods 4.2) for experimental complex cells and model complex cells. The mean values of C_v for experimental simple cells ($F_1/F_0 > 1$) and experimental complex cells ($F_1/F_0 < 1$) are 0.49 and 0.60 (Ringach et al., 2002), respectively. Figure 4.9 shows that values of C_v for experimental complex cells in Ringach et al. (2002) are mostly above 0.5 and have a wide coverage, while the values of C_v for model complex cells are mostly below 0.5 (mean 0.45, 0.42, and 0.45 for $N = 5$, 10, and 20, respectively). The mean C_v for model complex cells are even smaller than experimental simple cells, indicating model complex cells are selective to orientation.

Figure 4.10 shows the scatter plots of F_1/F_0 vs. half-bandwidth for experimental and model data. The bandwidths for models with different values of N are similar, with

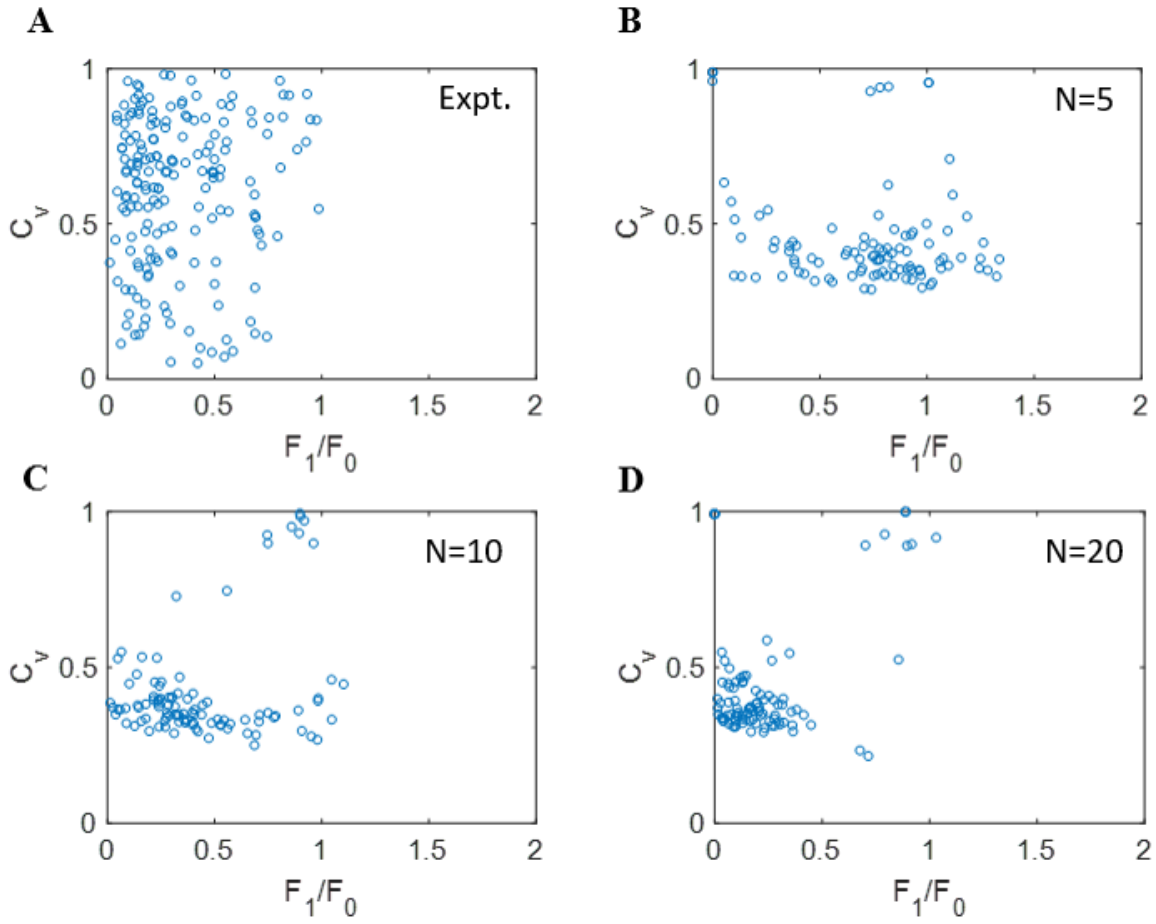

Figure 4.9: Scatter plots of F_1/F_0 vs. circular variance (C_V). (A) Experimental complex cells (Ringach et al., 2002). Model complex cells with (B) $N = 5$, (C) $N = 10$, and (D) $N = 20$.

means 27.3°, 25.2°, and 25.9° for $N = 5$, 10, and 20, respectively, which are similar to experimental complex cells that have mean half-bandwidth 30.6°.

Compared with the data for complex cells in an experimental study (Ringach et al., 2002) shown in Figure 4.8A, the histogram of F_1/F_0 for $N = 10$ (Figure 4.8C) is the most similar because it has more diversity for cells with $F_1/F_0 < 1$ compared with $N = 5$ and $N = 20$. The circular variances of most complex cells in the study of Ringach et al. (2002) are larger than model complex cells. For half-bandwidth of orientation tuning curves, model data is similar to the experimental study (Ringach et al., 2002) but with less variability. Given the better match to experimental data, the data set of $N = 10$ is used for further analysis in the following section. The discrepancies between model and experimental data are discussed in Discussion section 4.4.

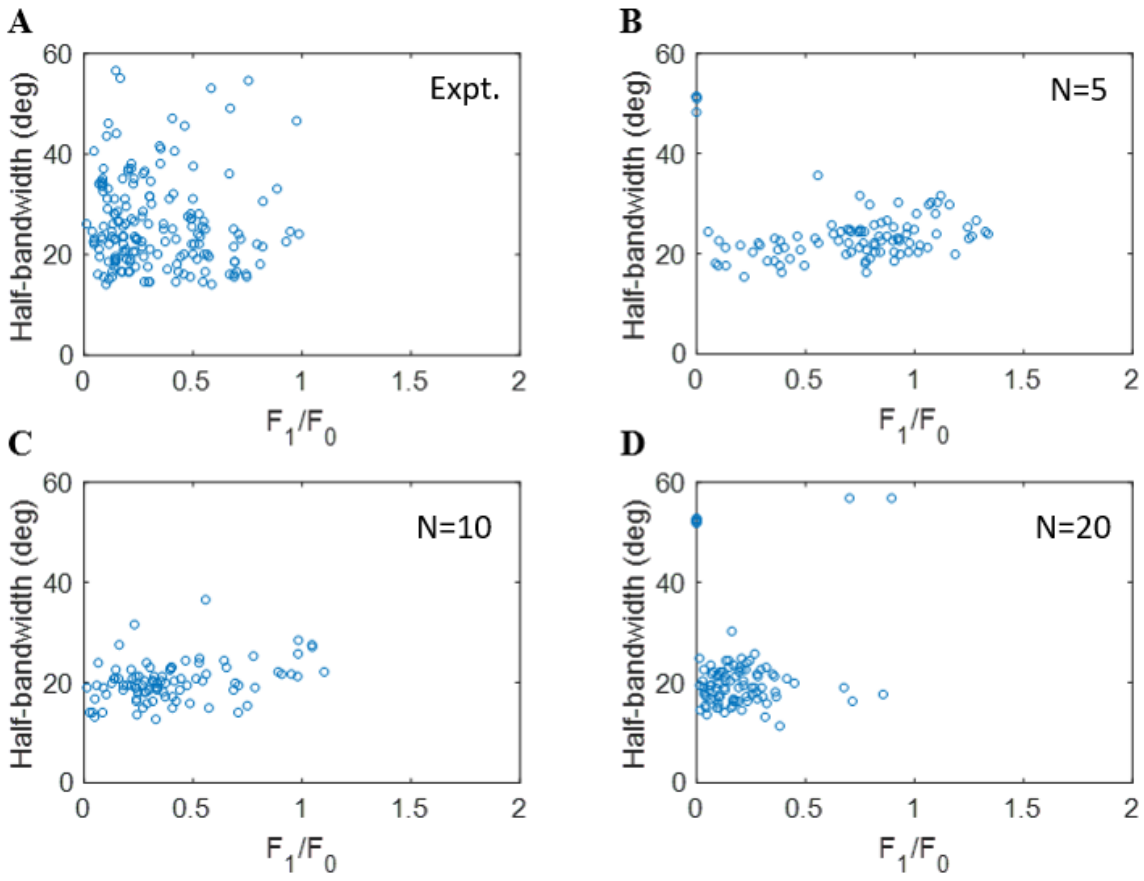

Figure 4.10: Scatter plots of F_1/F_0 vs. half-bandwidth. (A) Experimental complex cells (Ringach et al., 2002). (B) $N = 5$. (C) $N = 10$. (D) $N = 20$.

Examples of model complex cells

Some examples of complex cells are given in Figure 4.11 to demonstrate the diversity of model complex cells and the resemblance to experimental complex cells. In order to compare model complex cells with experimental complex cells (Almasi, 2017) investigated by NIM fitting, this method is also used to analyze model complex cells. After simulating the model responses to white noise, the model complex cell is fitted using NIM to find the spiking nonlinearity, filters, and upstream nonlinearity for the two filters (see Methods 4.2 for details).

Complex cell that is invariant to all spatial phases Figure 4.11 shows a complex cell that is invariant to all spatial phases. This complex cell has pronounced simple cell inputs with similar but not identical orientations (Figure 4.11A), and spatial phase tuning curves covering all phases, which leads to spatial phase invariance for the model complex cell with a very small ratio of $F_1/F_0 = 0.181$ (Figure 4.11C). The model complex cell is highly selective to one orientation as seen from Figure 4.11B. Figure 4.11D shows the spiking nonlinearity, filters, upstream nonlinearity for the two filters. The filters have almost the same orientation but different phases. In addition, both filters

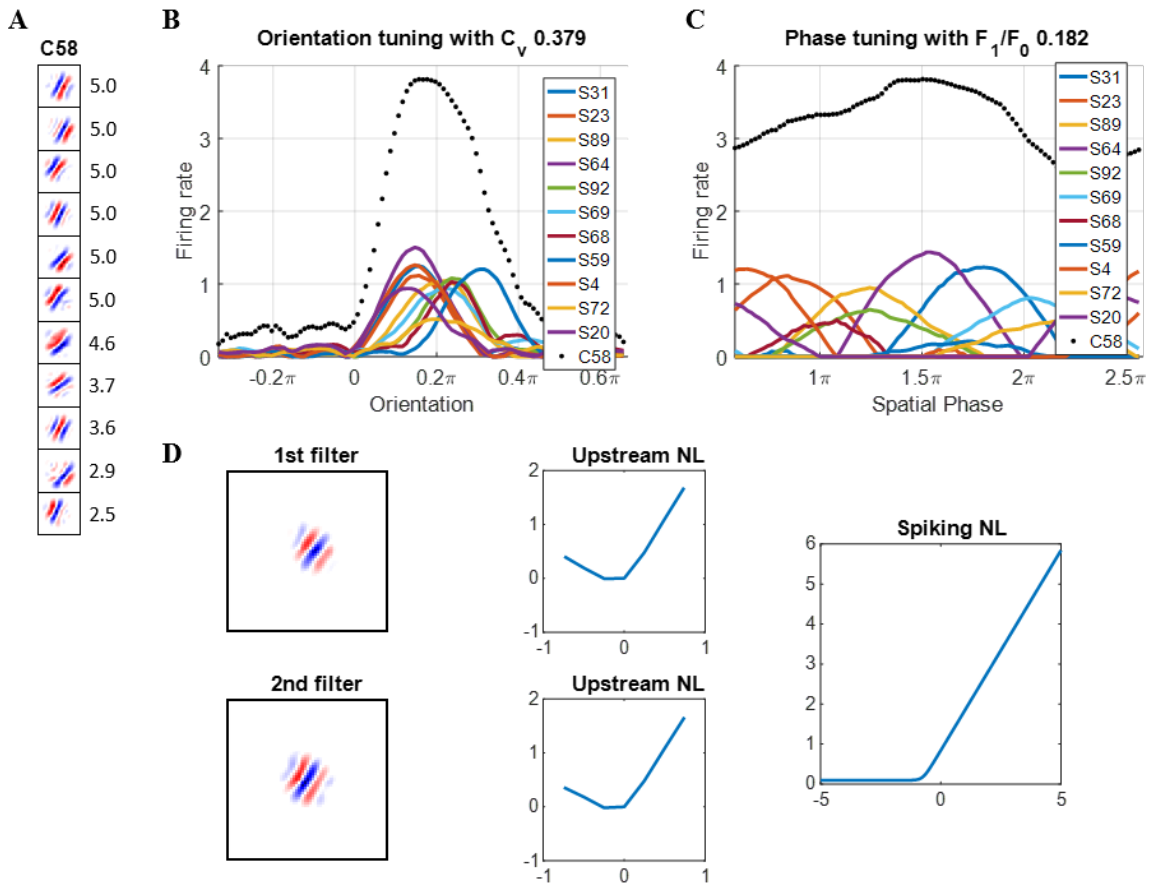

Figure 4.11: Complex cell C58. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C58. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Orientation tuning curve. (C) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dotted line is for complex cell C58. S represents simple cell and the following number is the index of the simple cell. (D) NIM fitting of complex cell C58. The filter, upstream nonlinearity (NL), and spiking NL are all fitted.

have a two-sided nonlinear function that is similar to the quadratic function in the energy model (Figure 3.1), which also explains spatial phase invariance in Figure 4.11C. The feature phase bandwidth, orientation breadth, spatial frequency breadth, and spatial phase breadth for cell C64 are 235°, 10°, 0.16 cycles/degree and 225°, respectively. Though the upstream nonlinearities of both filters are two-sided, they are not perfect quadratic functions as for the ideal energy model. Therefore, feature phase bandwidth and spatial phase breadth are large but not very close to 360°. The pooled simple cells have similar orientations, which leads to a small orientation breadth (10°). Qualitatively similar cells are observed when the NIM model is fitted to recorded responses in cat primary visual cortex (Almasi, 2017).

Complex cell that is invariant to a limited range of spatial phase Fig-

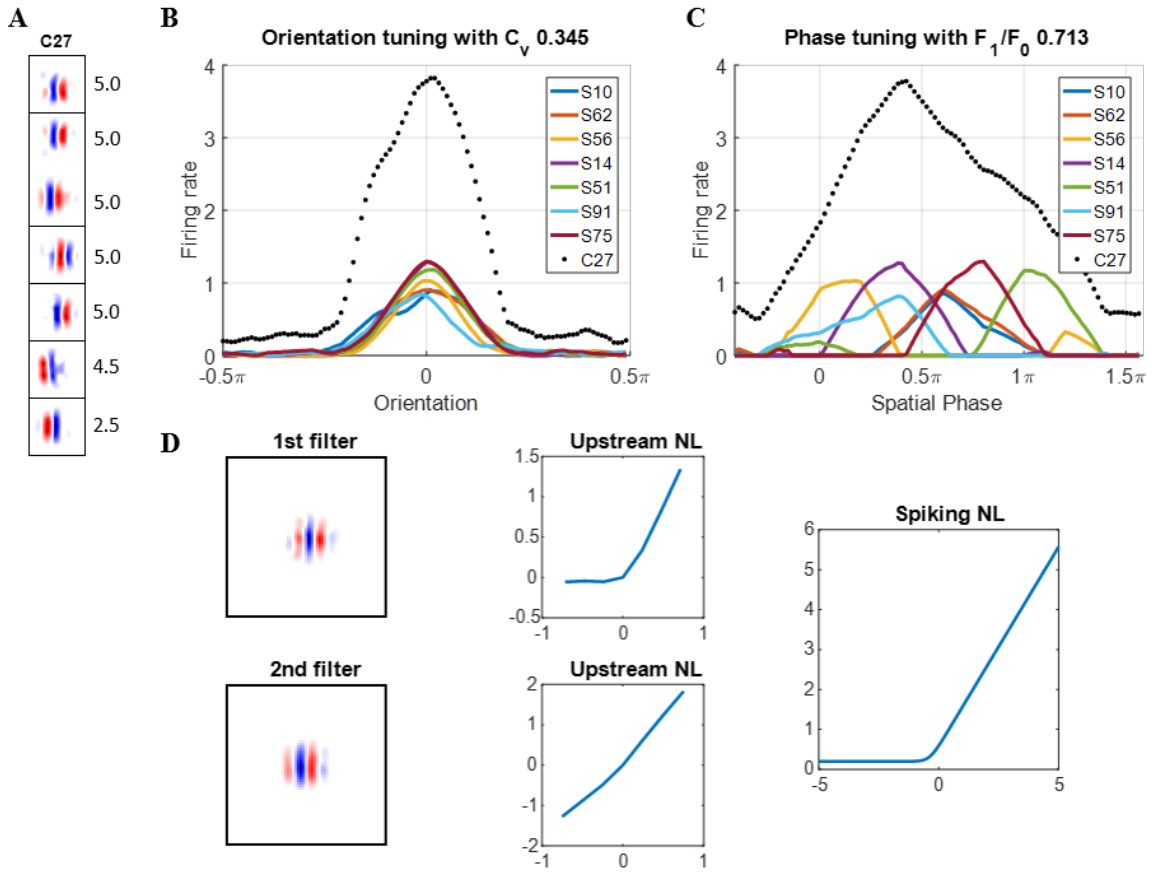

Figure 4.12: Complex cell C27. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C27. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Orientation tuning curve. (C) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dotted line is for complex cell C27. S represents simple cell and the following number is the index of the simple cell. (D) NIM fitting of complex cell C27. The filters, upstream nonlinearity (NL) and spiking NL are all fitted.

ure 4.12 shows a complex cell that has pronounced simple cell inputs with similar orientation. As seen from Figure 4.12C, the simple cells in the subspace have different spatial phase preferences, so the complex cell is more invariant to spatial phase than the pronounced simple cell inputs. However, this complex cell is only invariant to a limited region of spatial phase because the spatial phase tuning curves of the simple cells do not collectively cover the whole region of spatial phase. Figure 4.12B shows that the complex cell is selective to the vertical orientation. Figure 4.12D shows the spiking nonlinearity, filters, and the upstream nonlinearity for the two filters. The filters have the same orientation but different spatial phases. Compared to the complex cell in the previous example, the upstream nonlinearities have a different form, with the nonlinearity of the first filter being a one-side function, similar to a threshold-linear function, while the second nonlinearity is almost a linear function. These upstream nonlinearities are

far from a two-sided nonlinear function such as a quadratic function, which explains the partial invariance of the spatial phase tuning properties in Figure 4.12C. The feature phase bandwidth, orientation breadth, spatial frequency breadth, and spatial phase breadth for cell C27 are 165°, 5°, 0.14 cycles/degree, and 155°, respectively. The pooled simple cells have almost the same orientation, so the orientation breadth (5°) is small. Since the cell is only invariant to a limited region of spatial phase, the spatial phase breadth (155°) is moderate. Similar cells are observed when the NIM model is fitted to recorded responses from the experimental study of cat primary visual cortex (Almasi, 2017).

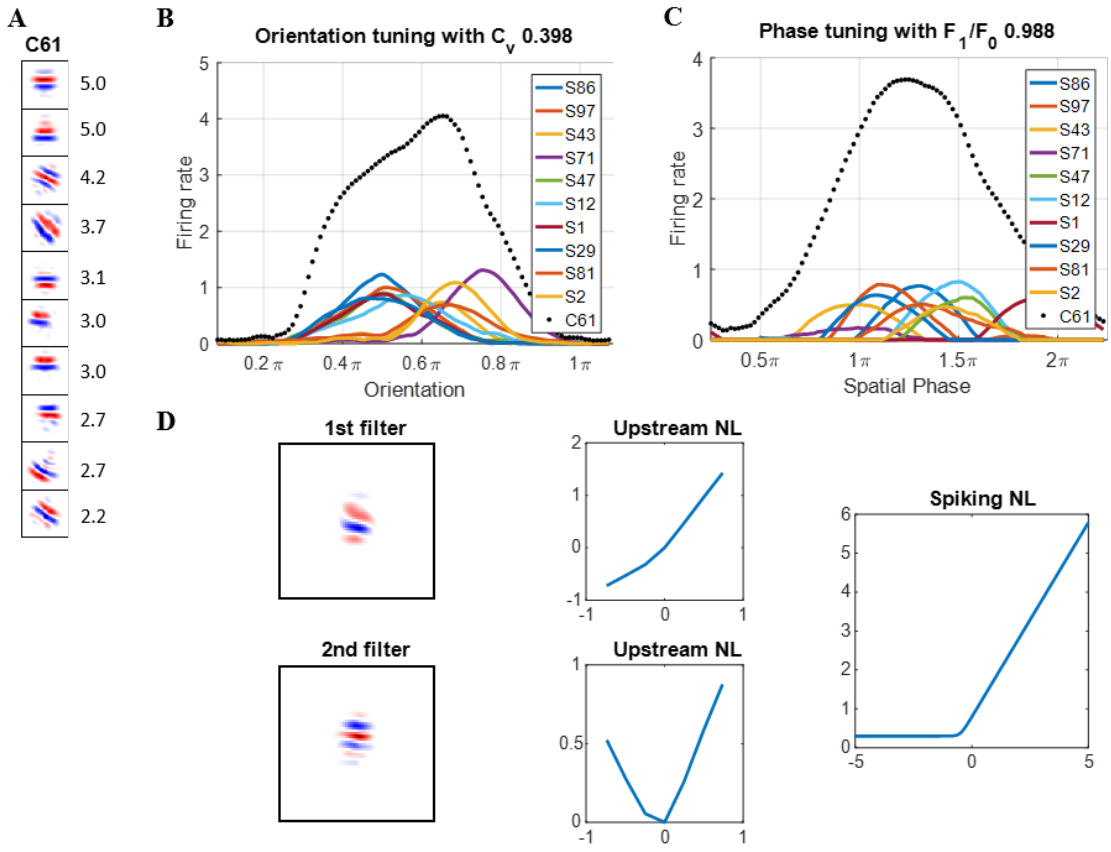

Figure 4.13: Complex cell C61. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C61. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Orientation tuning curve. (C) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dotted line is for complex cell C61. S represents simple cell and the following number is the index of the simple cell. (D) NIM fitting of complex cell C61. The filters, upstream nonlinearity (NL) and spiking NL are all fitted.

Complex cell that shows invariance to perturbations in orientation Figure 4.13 shows a model complex cell that has two major orientations in the subspace: one horizontal and another diagonal (Figure 4.13A). As a result, the orientation tuning curve (Figure 4.13B) has a wider bandwidth compared with other examples. Similar to complex

cell C1 (Figure 4.12), complex cell C61 is only invariant to a limited region of spatial phase because the spatial phase tuning curves of simple cells in the subspace only cover a subset of 2π region, as seen in Figure 4.13C. Figure 4.13D shows NIM fitting of complex cell C61, which shows that the first filter has a somewhat different orientation from the second filter and these two orientations are consistent with the two major orientations of simple cells in the subspace (Figure 4.13A). Filters with different orientations give the model complex cell nonlinear invariance to perturbations in orientation, as seen from a wider orientation tuning curve in Figure 4.13B. For the upstream nonlinearity, one is a two-sided function similar to the quadratic function while the other is almost a linear function, which makes complex cell C61 less invariant to spatial phase compared to complex cell C58 (Figure 4.11). The feature phase bandwidth, orientation breadth, spatial frequency breadth, and spatial phase breadth for cell C61 are 147°, 25°, 0.34 cycles/degree, and 80°, respectively. The pooled simple cells have two distinct orientations, so the orientation breadth (25°) is larger than the previous examples. Similar cells that have filters with different orientations are also observed when the NIM model is fitted to recorded responses of cat primary visual cortex (Almasi, 2017).

Complex cell that is invariant to orientation but not spatial phase Figure 4.14 shows an example cell that is invariant to orientation (Figure 4.14B) but not to spatial phase (Figure 4.14C). Figure 4.14A shows that simple cells in the subspace have various orientations from horizontal and diagonal to vertical orientations, even with non-oriented simple cells. Figure 4.14D shows that NIM fitting of this cell recovered one non-oriented filter and one ill-structured filter. The first filter with nearly one-sided nonlinearity can explain orientation invariance and polarity selectivity. The second ill-structured filter is caused by the overall effect of pooled simple cells with different orientations and shapes. The feature phase bandwidth, orientation breadth, spatial frequency breadth, and spatial phase breadth for cell C14 are 217°, 22.5°, 0.12 cycles/degree, and 135°, respectively.

Population statistics compared with experimental data

After showing some example model cells above, population statistics are analyzed using the four measurements described in the Methods section: feature-phase bandwidth, orientation breadth, spatial frequency breadth, and spatial phase breadth. Comparisons between model and experimental data (Almasi, 2017) are shown in Figure 4.15.

Figure 4.15A shows that model data covers similar regions compared with experimental data. The distributions of model and experimental data are similar except that model data has a higher percentage of cells with feature phase bandwidth close to 180 degrees.

Figure 4.15B shows that most cells for both model and experimental data cover a small region of orientation, but experimental complex cells have a somewhat broader

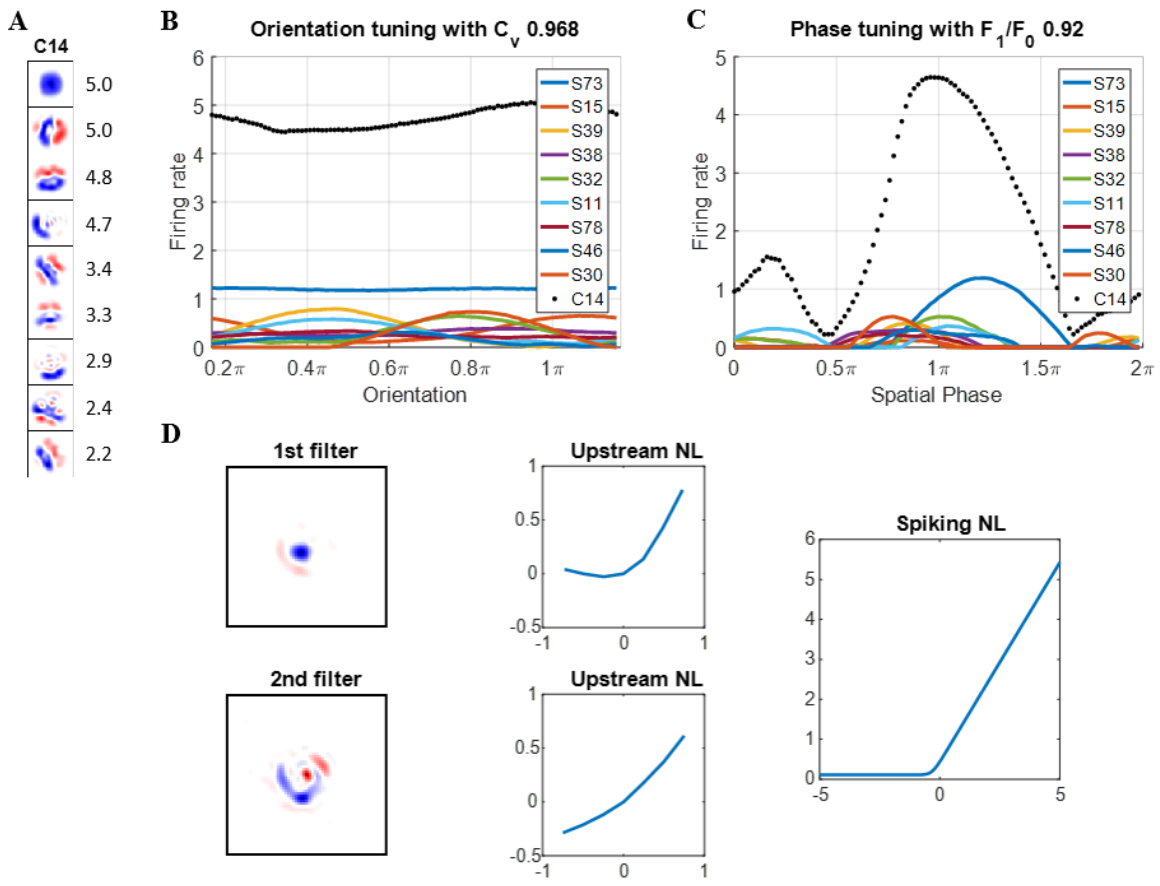

Figure 4.14: Complex cell C14. (A) Each block in subplot A is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace and the number on the right is the feedforward weight connected with complex cell C14. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. (B) Orientation tuning curve. (C) Spatial phase tuning curves. Solid lines are for simple cells in the subspace. Dotted line is for complex cell C14. S represents simple cell and the following number is the index of the simple cell. (D) NIM fitting of complex cell C14. The filter, upstream nonlinearity (NL) and spiking NL are all fitted.

distribution of orientation breaths than the model complex cells.

Figure 4.15C shows that model data has similar coverage of spatial frequencies as experimental data except that the model data has more complex cells with small spatial frequency breadth. It is important to mention that the spatial frequency of the model was calculated based on the assumption that the input image is equivalent to 10 degrees of the visual field. Therefore, different sizes of the visual field will scale the spatial frequency, which might bring the model data closer to the experimental data.

Figure 4.15D shows that both model and experimental data cover a wide range of spatial phase except that the model data has more complex cells with spatial breadth around 180 degrees.

Overall, the model can account for the diversity of complex cells found in a recent

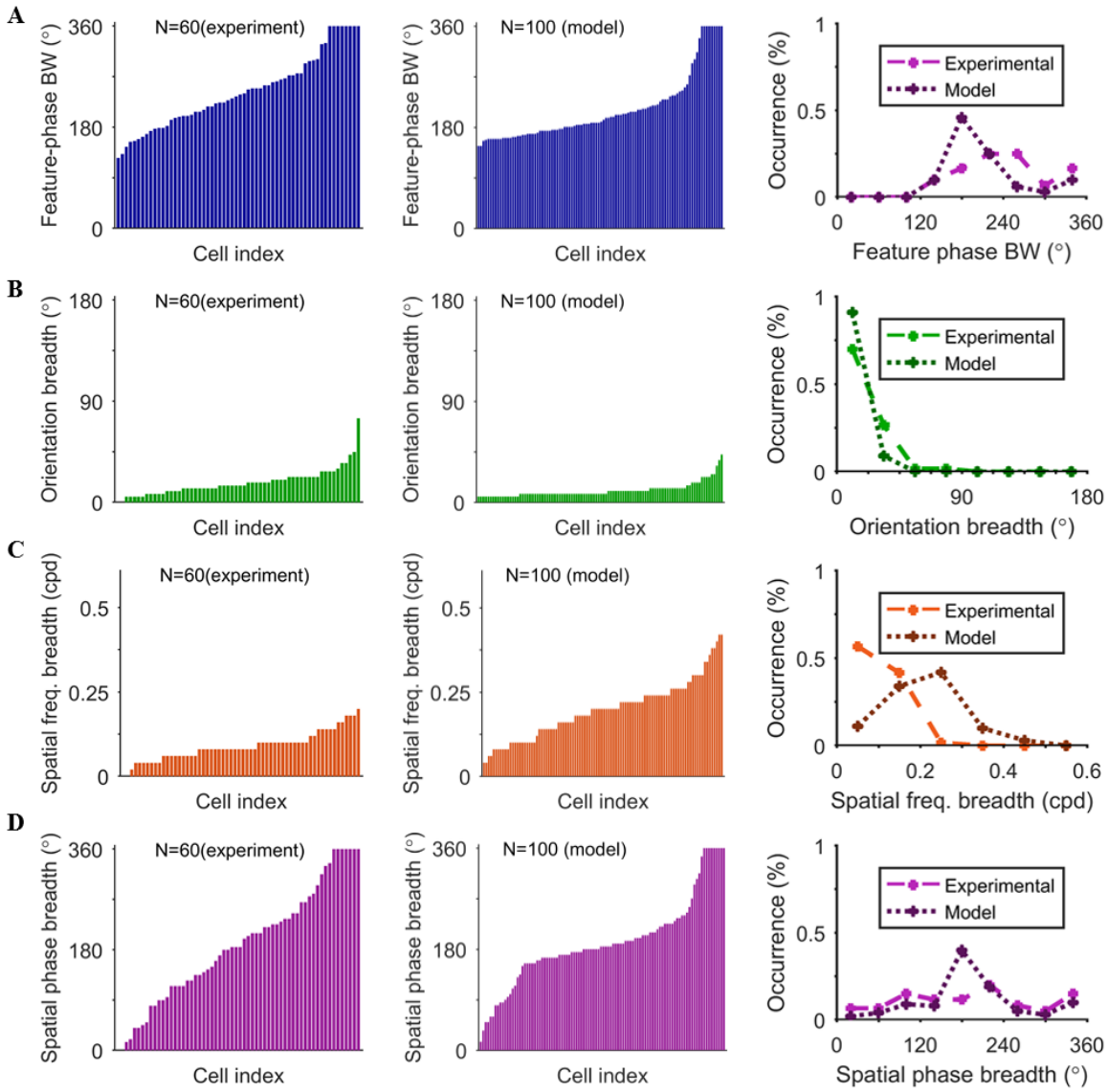

Figure 4.15: Comparison between experimental data (Almasi, 2017) (1st column) and model data (2nd column). (A) Feature phase bandwidth (degrees). (B) Orientation breadth (degrees). (C) Spatial frequency breadth (circles per degree). (D) Spatial phase breadth (degrees).

experimental study (Almasi, 2017). Despite some discrepancies in the histograms of these four measurements, the model can capture the trend of the distributions.

4.4 Discussion

Both the modified BCM and modified NBCM plasticity rules can learn complex cells that are spatial phase invariant and orientation selective, but the modified BCM rule failed to learn different complex cells compared with the modified NBCM rule. The similarities between the distributions of experimental and model data (learned by the modified NBCM

rule) suggests that the modified NBCM rule can explain complex cell properties well and that complex cells can be learned in a biologically plausible neural network.

Normalization in this model, defined previously as

$$r^{\mathrm{C}}_{j,N} = \frac{\beta r^{\mathrm{C}}_j}{\alpha + \sqrt{\sum_k (r^{\mathrm{C}}_k)^2}}, \tag{4.11}$$

is important for the model to learn different complex cells because it introduces competition to the network. Without normalization, model complex cells tend to be similar, without the diversity seen in experimental data. Normalization of responses in Eq. 4.8 was first proposed by Heeger (1992) and is suggested to be a canonical neural computation, for which there is increasing evidence (Carandini and Heeger, 2012). Rubin et al. (2015) suggested that normalization can be implemented by the cortex using the stabilized supralinear network that employs recurrent connections. The results presented in this study indicate that the model can learn complex cells using a hierarchical structure, but this does not rule out the importance of the recurrent structure because recurrent connections might be important to implement normalization in the neural circuits.

As discussed in Chapter 3, efficient coding, implemented as a sparse coding model, can learn the subspace of complex cells by finding an efficient representation of natural stimuli with temporal information. However, the introduced competition suppresses responses of model complex cells such that they behave like simple cells that are very selective to spatial phase. However, the model proposed in this chapter can learn receptive properties of complex cells. The modified BCM rule can learn useful representations and the normalization introduces competition to the network. In terms of competition, normalization is much softer than efficient coding because efficient coding pushes the activity of many cells to zero while normalization only adjusts the levels of activities. Therefore, the soft competition introduced by normalization helps the model learn different complex cells.

The model proposed in this chapter can pool simple cells into the subspace of complex cells and can account for the diversity of complex cells in a recent experimental study Almasi (2017). However, although the main features of the model agree qualitatively with the experimental data, some discrepancies between model and experimental data do exist and remain to be explored.

The experimental data shows more diversity in orientation bandwidths and circular variance than model data. As discussed in Willmore et al. (2012), the normalization constants in NBCM will affect the results. Furthermore, the values of learning rates and other parameters might also lead to different result sets. Another reason for differences in orientation bandwidths and circular variance between the model and experimental data might be related to the visual stimuli. In the experimental study that measured

orientation selectivity in macaque V1 (Ringach et al., 2002), drifting sinusoidal gratings were used as the visual stimuli. The cortical cells might have some temporal dynamics of neuronal responses related to the drifting gratings. However, when calculating the spatial phase tuning properties of model cells, steady state responses to each spatial phase of the drifting gratings were used. In addition, the current model does not incorporate the temporal dynamics of cells. The investigation of the model that incorporates temporal dynamics and compares different values of parameters is left for future research.

The model can account for the diversity of complex cells in the cat visual cortex reported by Almasi, but some differences do exist in the distributions of the population statistics. This may be because model cells in this chapter are only a subset of the rich repertoire of real cortical cells, and choices of free parameters in the model might also lead to different results. Furthermore, filters recovered by NIM fitting in the experimental study of Almasi (2017) tended to have fewer stripes than the model. A possible reason is that Almasi's study used data from cat primary visual cortex, while receptive fields of monkey V1 tend to have more stripes (Ringach, 2002).

Another group recently presented a modeling work of complex cells based on predictive coding (Franciosini et al., 2019). The difference from the model proposed here is that their model uses symmetric connections and max pooling to learn a topographical map of simple cells such that model complex cells pool local simple cells that are similar to achieve phase invariance.

4.5 Conclusion

In this chapter, a plausible learning model with a hierarchical structure of complex cells is proposed. After training, the model pools simple cells into the subspaces that can account for the diversity of receptive field properties of complex cells in experimental studies. The close match of the main features of the population statistics between model and experimental data (Almasi, 2017) provides strong support for the hierarchical structure in the cortex. However, recurrent connections are also expected to play an important role through their ability to implement normalization, which helps the model learn different complex cells.

Chapter 5

Conclusion

5.1 Contributions

In this thesis I investigated how the primary visual cortex processes visual information by designing computational models that incorporate biological constraints that have previously been largely neglected, such as a local learning rule and Dale's Law. This research provides insights for better understanding how the brain processes visual information and potentially uncovers some of the more general principles of how the brain functions, particularly in sensory areas.

I proposed biologically-constrained models for simple cells and complex cells. After the models were trained on natural images, I compared the properties of model receptive field (RF) properties with experimental data. Three major contributions were made in this thesis:

- I developed a computational model of simple cells based on efficient coding that incorporates biological constraints such as local learning rule and Dale's law, and showed that the model can account for various experimental phenomena such as the separation of ON and OFF sub-regions of simple cell receptive fields, push-pull effect of simple cells, phase-reversed cortico-thalamic feedback, contrast invariance of simple cells, and diversity of simple cell receptive fields. The model suggests that sparse coding can be implemented using simple neural circuits with biologically plausible architecture.

- I investigated the efficient coding model of complex cells with biological constraints. When biological constraints are incorporated, results demonstrate that visual input with temporal information is necessary for the model complex cell to learn the subspace with various spatial phase preferences. However, the competition introduced by the principle of efficient coding suppresses responses of model complex cells such that they fail to account for spatial phase invariance of experimental complex cells.

- I developed a model of complex cells based on Bienenstock, Cooper and Munro (BCM) plasticity rule while incorporating biological constraints. The results here demonstrate that this model can learn complex cells and that they have similar properties to those found in a recent experimental study of cat V1. Results showed that model complex cells can capture the population statistics of orientation, spatial phase and spatial frequency for experimental complex cells.

Computational models of primary visual cortex proposed in this thesis bring neural models closer in their response properties to those in the real biological system. This is crucial to understanding how the brain works because (1) previously proposed artificial models ignore many constraints that are important for the biological system, (2) the biologically plausible model proposed here is a more realistic model that can investigate various aspects of the brain and explain more experimental phenomena, and (3) the model developed here provides a framework upon which future studies can be built to uncover more detailed properties of the brain.

Properties of V1 cells in the model are a direct result of learned connections between populations, suggesting that plasticity plays a key role in building the network structure that underlies the observed neuronal activity. The learning ability of the model is very important because the neural circuits of the brain are not pre-wired and brain functions are highly dependent upon the plasticity of synaptic connections.

The hierarchical structure in the model is crucial to learning RF properties of simple and complex cells. Recurrent connections are also important to implement the model in neural circuits, which may provide insights into more general principles of how the brain functions. The biological neural circuits of the brain have a clear hierarchical structure, where more complicated features are processed in higher areas, but the brain also employs a large amount of lateral and recurrent connections. Models proposed in this thesis suggest that the framework with the hierarchical structure and recurrent connections is a potential approach to investigate many higher level brain functions.

Finally, the work presented in this thesis helps us to better understand vision, with implications for deep learning techniques in artificial intelligence.

5.2 Future work

The research presented in this thesis has some limitations that could be addressed in future work. Chapter 2 provides a biologically plausible model of LGN-V1 pathways, but the neuron model is rate-based and does not incorporate spiking dynamics of the sort required for a more detailed neuronal model. Although, there exist different types of excitatory and inhibitory neurons in V1, the current model does not incorporate neuron types. How the inhibitory connections between LGN and simple cells can be implemented

in the model is not described in this thesis in detail. Incorporating the above features into a model will lead to more specific microcircuits that are closer to the real biological system and may help with the understanding of brain function in greater detail.

Secondly, the model of complex cells takes natural images that have temporal information as the input, which is implemented using the average activity of simple cells in response to image patches in nearby regions. However, real cortical cells have a mechanism of responding to temporal stimuli. Therefore, a more detailed model with the temporal dynamics of simple and complex cells would better describe the properties of V1 neurons.

Furthermore, the learning rule of model complex cells utilizes response normalization. Normalization is found to be an important principle for many visual areas, but the model proposed in this thesis does not have a specific realisation of normalization. Though some neural circuits are proposed to implement normalization, it is still not clear what neural circuit carries out this function in the real biological system. A more biologically plausible model that combines learning and the neural circuits of normalization will likely provide more insights into the microcircuits of the brain and potentially explain various experimental phenomena.

Additionally, the work presented in this thesis can be extended to investigate the effects of different weight normalization techniques and different choices of model parameters. In addition, it would be interesting to explore the relationship between the non-linearities introduced in the model by sparsity of neuronal activity.

Finally, the work presented here focused on area V1. V1 is only the first cortical area that processes visual information and there are numerous higher areas in the visual pathway, such as V2 and V4, that process more complicated visual features. A biologically plausible model that explains how higher areas of the brain process visual information is clearly an important area of ongoing and future research.

In summary, future work remains to be done to build a more detailed model of brain networks that incorporates temporal dynamics, different types of neurons and neural circuits for implementing normalization. Hopefully, the work presented in this thesis forms part of the basis upon which future studies can extend the analysis to investigate higher areas in the cortex, in order to help unveil the secrets of the brain.

Appendix A

Computing the F1/F0 ratio for a sequence of cell activities in response to sinusoidal gratings

A.1 Theoretical Procedure

The temporal stimulus used here is drifting sinusoidal grating in a circular window. The process of measuring F_1/F_0 of a cell is as follows:

1. Find the preferred sinusoidal grating for this cell.
   Search for the sinusoidal grating with certain orientation, phase, spatial frequency, center and the radius of the circle that gives the maximum firing rate. Assume the preferred sinusoidal grating has the following form:

   $$g = \sin\left(2\pi f_x[(x - x_0)\cos\theta + (y - y_0)\sin\theta] + \phi\right) \tag{A.1}$$

   where f_x is the spatial frequency, θ is the orientation, ϕ is the phase and (x_0, y_0) is the center of the grating. We denote R as the optimal radius (the size) of the sinusoidal grating in a circular window.

2. Display the drifting sinusoidal grating.
   The drifting sinusoidal grating has the following form:

   $$g(t) = \sin\left(2\pi f_x[(x - x_0)\cos\theta + (y - y_0)\sin\theta] + \phi - 2\pi f_t t\right) \tag{A.2}$$

   where f_t is the temporal frequency.

3. Record the firing rates $r(t)$ of the cell when drifting sinusoidal grating is presented.

4. F_1/F_0 is just the ratio between the amplitude of the first harmonic of $r(t)$ and the mean spike rate.
   $r(t)$ can be written into the sum of harmonic series:

$$\begin{aligned} r(t) &= A_0 + \sum_{k=1}^{\infty} A_k \cos(2\pi k f_t t + \phi_k) \\ &= A_0 + \sum_{k=1}^{\infty} a_k \cos(2\pi k f_t t) + b_k \sin(2\pi k f_t t), \end{aligned} \tag{A.3}$$

   where $A_k = \sqrt{a_k^2 + b_k^2}$ and $\phi_k = \arctan(b_k/a_k)$.
   Then,

$$F_1/F0 = \frac{|A_1|}{|A_0|}. \tag{A.4}$$

A.2 Real Application in Discrete Time Domain

The procedure in mentioned above is always carried out in discrete time, so we explain how the theoretical procedure shall be used in real application to measure F_1/F_0.

First we denote f_s as the sampling frequency.

1. Find the preferred sinusoidal grating for this cell.

2. Display the drifting sinusoidal grating in discrete time domain.
   The sampled discrete drifting sinusoidal grating has the following form:

$$g[n] = g(\frac{n}{f_s}) = \sin\left(2\pi f_x[(x - x_0)\cos\theta + (y - y_0)\sin\theta] + \phi - 2\pi\frac{f_t}{f_s}n\right) \tag{A.5}$$

3. Record the firing rates $r[n]$ of the cell when drifting sinusoidal grating is presented.

4. Computing F_1/F_0.
   Perform discrete Fourier transform (DFT) on $r[n]$. Denote f_0 as the magnitude of DFT at DC frequency. Denote f_1 as the magnitude of DFT at (or near) frequency of f_t.
   Then

$$F_1/F_0 = 2\frac{f_1}{f_0}. \tag{A.6}$$

A.3 Equation A.6 comes from Equation A.4

A.3.1 Compute F1/F0 from harmonic series

From Eq. (A.3), $r[n]$ can be written into

$$\begin{aligned} r[n] &= A_0 + \sum_{k=1}^{\infty} A_k \cos\left(2\pi k \frac{f_t}{f_s} n + \phi_k\right) \\ &= A_0 + \sum_{k=1}^{\infty} a_k \cos\left(2\pi k \frac{f_t}{f_s} n\right) + b_k \sin\left(2\pi k \frac{f_t}{f_s} n\right). \end{aligned} \tag{A.7}$$

Denote $f_d = f_t/f_s$. Assume that we have sampled N data points for a period T $(1/f_t)$. Then, we can get that

$$\begin{aligned} &\sum_{n=0}^{N-1} \cos(2\pi k f_d n) = 0 \\ &\sum_{n=0}^{N-1} \sin(2\pi k f_d n) = 0, \text{ for k larger than 0} \end{aligned} \tag{A.8}$$

Therefore,

$$\begin{aligned} \sum_{n=0}^{N-1} r[n] &= \sum_{n=0}^{N-1} A_0 + \sum_{k=1}^{\infty} \sum_{n=0}^{N-1} a_k \cos(2\pi k f_d n) + \sum_{k=1}^{\infty} \sum_{n=0}^{N-1} b_k \sin(2\pi k f_d n) \\ &= N A_0 \end{aligned} \tag{A.9}$$

i.e.

$$A_0 = \sum_{n=0}^{N-1} r[n]/N.$$

In addition,

$$\begin{aligned} \sum_{n=0}^{N-1} r[n] \cos(2\pi f_d n) = A_0 &\sum_{n=0}^{N-1} \cos(2\pi f_d n) + \\ &\sum_{k=1}^{\infty} \sum_{n=0}^{N-1} a_k \cos(2\pi k f_d n) \cos(2\pi f_d n) + \\ &\sum_{k=1}^{\infty} \sum_{n=0}^{N-1} b_k \sin(2\pi k f_d n) \cos(2\pi f_d n) \\ = &\sum_{n=0}^{N-1} 0.5 * a_1 = N * a_1/2, \end{aligned} \tag{A.10}$$

i.e.

$$a_1 = 2/N \sum_{n=0}^{N-1} r[n] \cos(2\pi f_d n).$$

Similarly, we can get

$$b_1 = 2/N \sum_{n=0}^{N-1} r[n] \sin(2\pi f_d n).$$

Therefore,

$$\begin{aligned} F_1/F_0 = \frac{|A_1|}{|A_0|} &= \frac{\sqrt{(a_1^2 + b_1^2)}}{A1} \\ &= \frac{\frac{2}{N}\sqrt{\left(\sum_{n=0}^{N-1} r[n]\cos(2\pi f_d n)\right)^2 + \left(\sum_{n=0}^{N-1} r[n]\sin(2\pi f_d n)\right)^2}}{\frac{1}{N}\left|\sum_{n=0}^{N-1} r[n]\right|} \\ &= \frac{2\sqrt{\left(\sum_{n=0}^{N-1} r[n]\cos(2\pi f_d n)\right)^2 + \left(\sum_{n=0}^{N-1} r[n]\sin(2\pi f_d n)\right)^2}}{\left|\sum_{n=0}^{N-1} r[n]\right|} \end{aligned} \tag{A.11}$$

A.3.2 Compute F1/F0 from discrete Fourier transform

Now we take a M-point DFT of $r[n]$, i.e.

$$\begin{aligned} R[m] &= \sum_{n=0}^{N-1} r[n] e^{-j\frac{2\pi mn}{M}} \\ R[m] &= \sum_{n=0}^{N-1} r[n] \cos\left(\frac{2\pi mn}{M}\right) - j \sum_{n=0}^{N-1} r[n] \sin\left(\frac{2\pi mn}{M}\right) \end{aligned} \tag{A.12}$$

for $m = 0, 1, \cdots, M-1$.

We can get $R[0] = \sum_{n=0}^{N-1} r[n]$. The magnitude response at DC is

$$f_0 = |R[0]| = \left|\sum_{n=0}^{N-1} r[n]\right|.$$

Assume that there exist an integer p such that $p/M = f_d = f_t/f_s$. Then

$$\begin{aligned} R[p] &= \sum_{n=0}^{N-1} r[n] \cos\left(\frac{2\pi pn}{M}\right) - j \sum_{n=0}^{N-1} r[n] \sin\left(\frac{2\pi mn}{M}\right) \\ &= \sum_{n=0}^{N-1} r[n] \cos\left(2\pi f_d n\right) - j \sum_{n=0}^{N-1} r[n] \sin\left(2\pi f_d n\right) \end{aligned} \tag{A.13}$$

The magnitude response at the frequency of f_t is

$$|R[p]| = \sqrt{\left(\sum_{n=0}^{N-1} r[n]\cos(2\pi f_d n)\right)^2 + \left(\sum_{n=0}^{N-1} r[n]\sin(2\pi f_d n)\right)^2}$$

.

Therefore,

$$\begin{aligned}\frac{2f_1}{f_0} &= \frac{2|R[p]|}{|R[0]|} \\ &= \frac{2\sqrt{\left(\sum_{n=0}^{N-1} r[n]\cos(2\pi f_d n)\right)^2 + \left(\sum_{n=0}^{N-1} r[n]\sin(2\pi f_d n)\right)^2}}{\left|\sum_{n=0}^{N-1} r[n]\right|} \\ &= \frac{2|A_1|}{|A_0|},\end{aligned} \tag{A.14}$$

which means Eq. A.6 computes the right F_1/F_0 ratio. If we take $M = 1000$ and $f_s = 100f_t$, we will have $p = 100$ and

$$F_1/F_0 = 2\frac{|R[p]|}{|R[0]|}, \tag{A.15}$$

which is Eq. 3.8.

Appendix B

Full subspaces of model complex cells

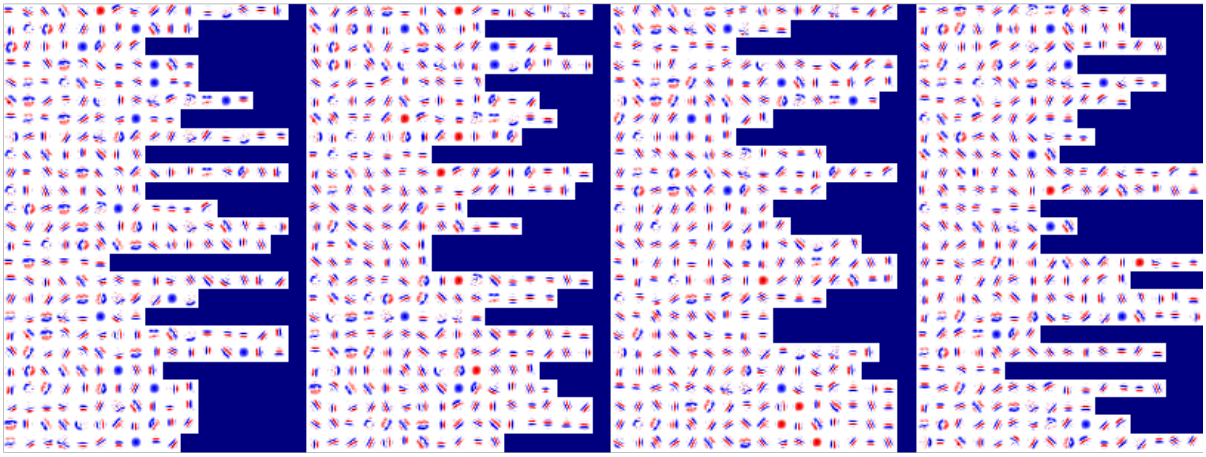

Figure B.1: Subspace of all 100 model complex cells using efficient coding with $\lambda_C = 0.1$ on static natural images (in Chapter 3). C represents complex cell and the followed number is the index of the complex cell. Each block is a 16×16 synaptic field (defined in Eq. 3.10) of simple cells that have a feedforward weight larger than 0.1 in the subspace of a complex cell. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. The figure only shows up to 16 subunits.

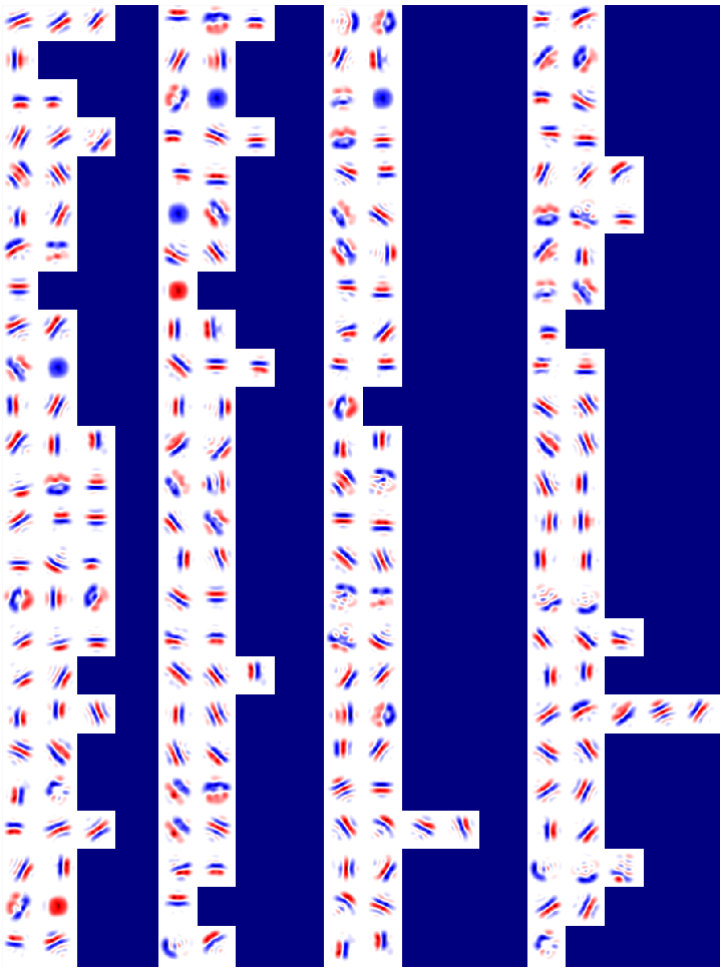

Figure B.2: Subspace of all 100 model complex cells using efficient coding with $\lambda_C = 0$ on static natural images (in Chapter 3). C represents complex cell and the followed number is the index of the complex cell. Each block is a 16×16 synaptic field (defined in Eq. 3.10) of simple cells that have a feedforward weight larger than 0.1 in the subspace of a complex cell. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. The maximum number of subunits is 5 and thus the figure only shows up to 5 subunits.

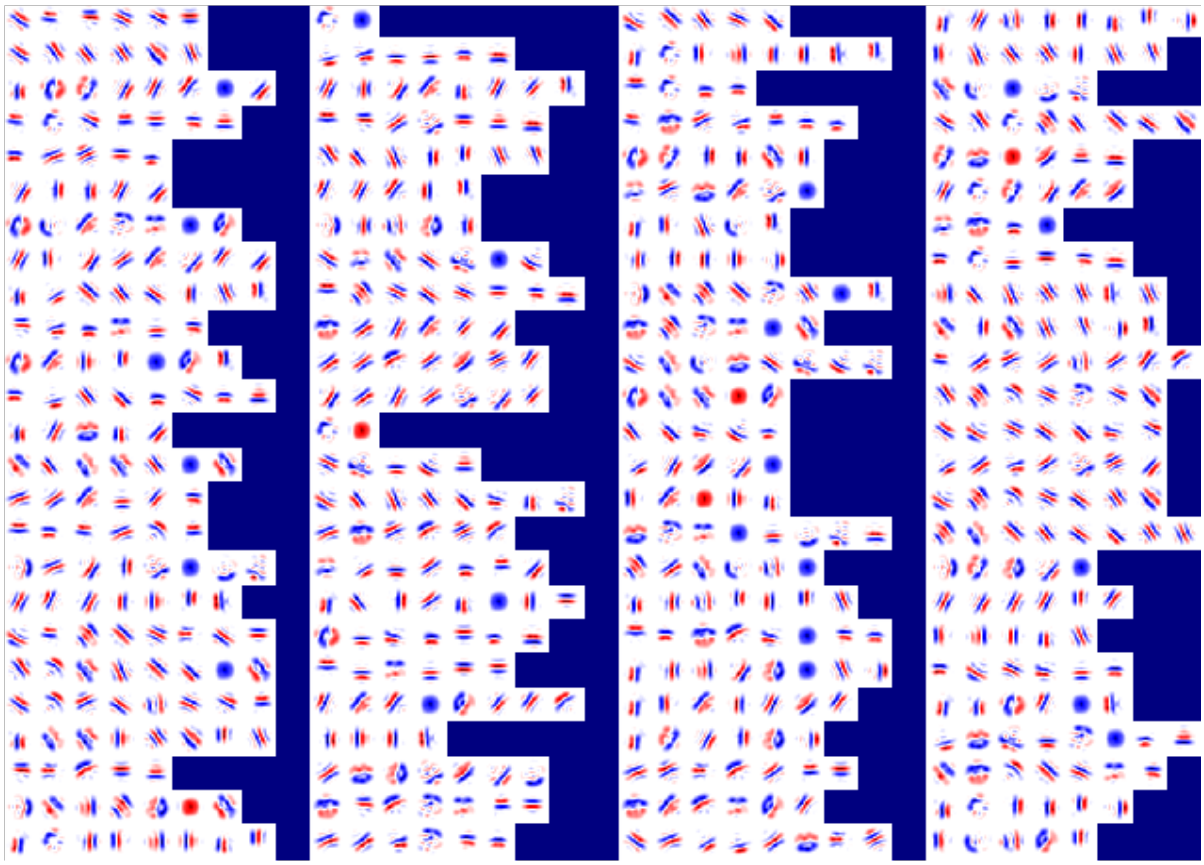

Figure B.3: Subspace of all 100 model complex cells using efficient coding with $\lambda_C = 0$ on natural image with jitters (in Chapter 3). C represents complex cell and the following number is the index of the complex cell. Each block is a 16×16 synaptic field (defined in Eq. 3.10) for simple cells in the subspace of a complex cell. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure. The figure only shows up to 8 subunits.

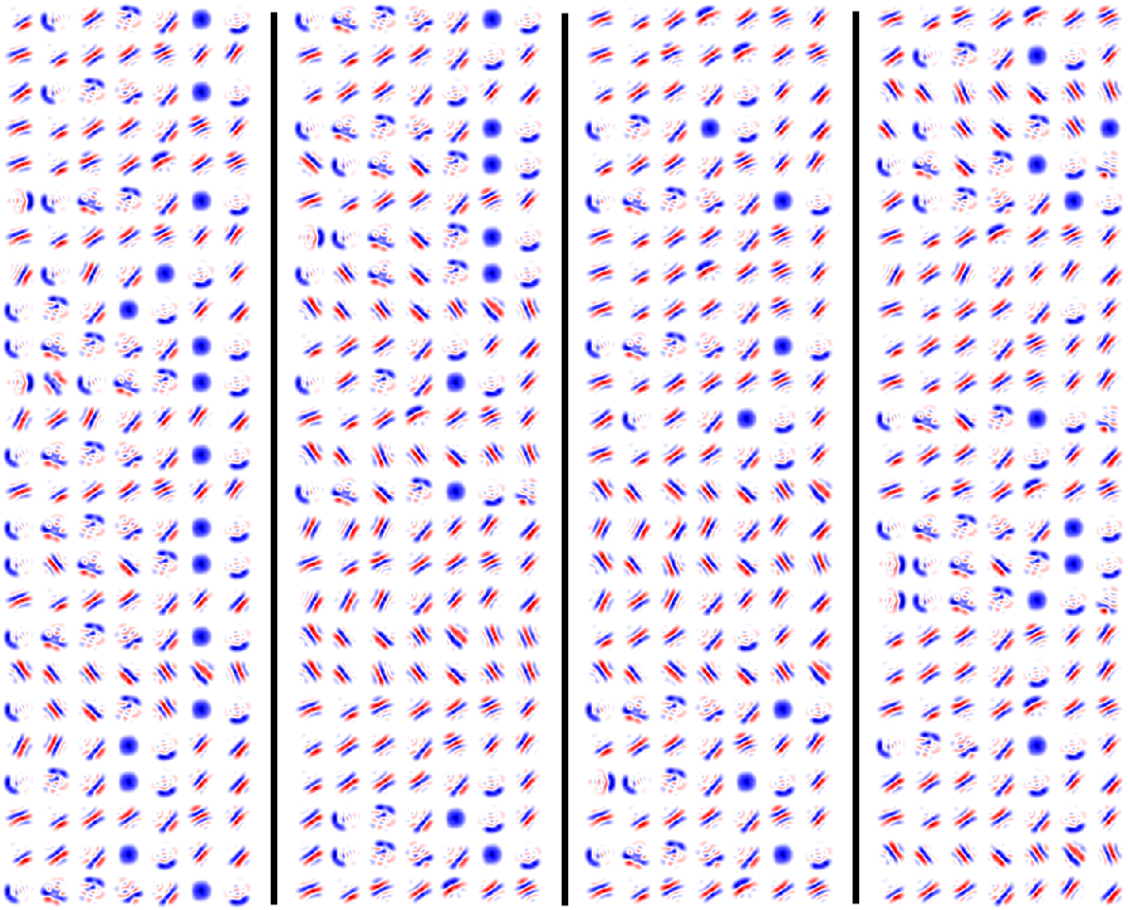

Figure B.4: Subspace of all 100 model complex cells based on modified BCM rule (in Chapter 4). There are 25 rows and each row has the subspace for four complex cells. The simple cells in each subspace have connection weight larger than 2. If a complex cell has more than seven simple cells in the subspace, only the top seven significant simple cells are displayed. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure.

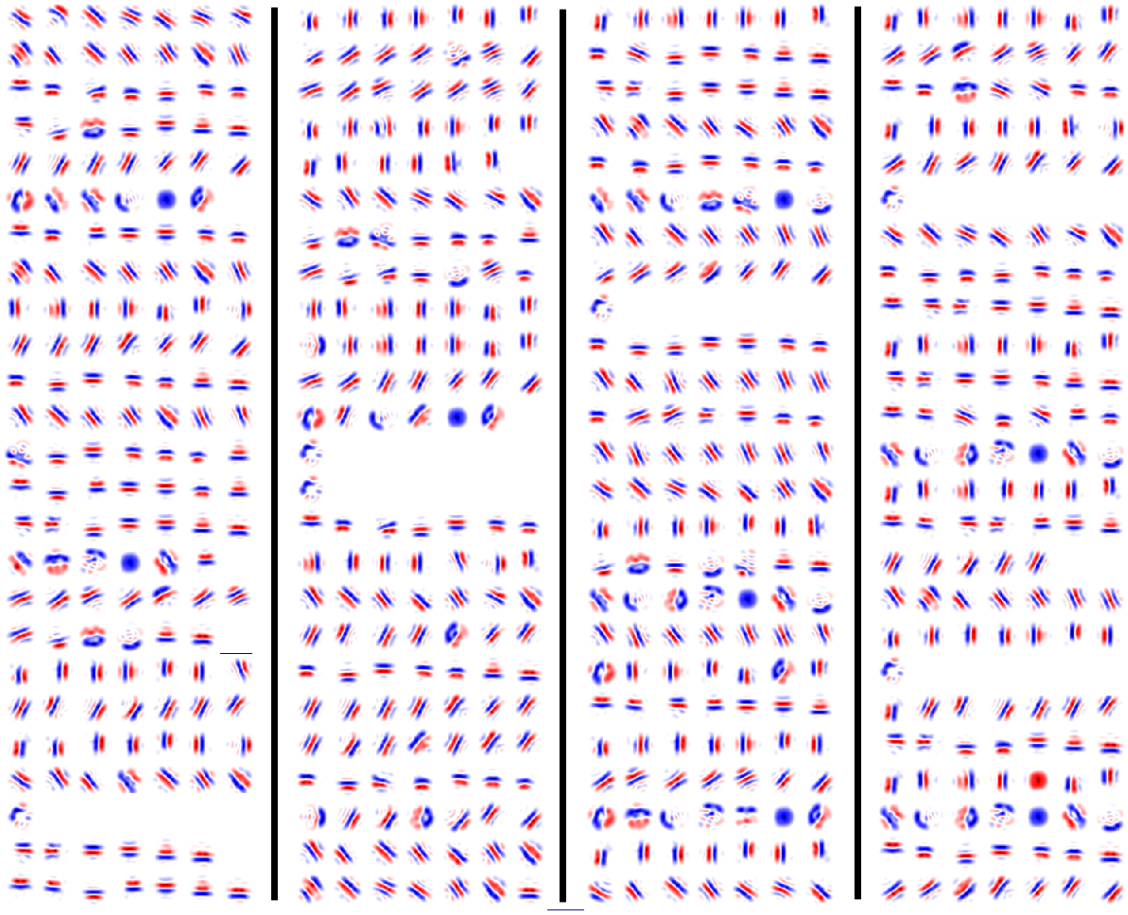

Figure B.5: Subspace of all 100 model complex cells based on modified NBCM rule (in Chapter 4). There are 25 rows and each row has the subspace for four complex cells. The simple cells in each subspace have connection weight larger than 2. If a complex cell has more than seven simple cells in the subspace, only the top seven significant simple cells are displayed. Values in each block are normalized to the range $[-1\ 1]$ when plotting the figure.

Bibliography

Abbott, L. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain Res. Bull.*, 50(5-6):303–304.

Adelson, E. and Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299.

Ahmed, B., Anderson, J., Douglas, R., Martin, K., and Nelson, J. (1994). Polyneuronal innervation of spiny stellate neurons in cat visual cortex. *J. Comp. Neurol.*, 341(1):39–49.

Alitto, H. and Usrey, W. (2004). Influence of contrast on orientation and temporal frequency tuning in ferret primary visual cortex. *J. Neurophysiol.*, 91(6):2797–2808.

Almasi, A. (2017). *An investigation of Spatial receptive fields of complex cells in the primary visual cortex*. PhD thesis, The University of Melbourne.

Antolik, J. and Bednar, J. (2011). Development of maps of simple and complex cells in the primary visual cortex. *Front. Comput. Neurosci.*, 5:17.

Atick, J. and Redlich, A. (1992). What does the retina know about natural scenes? *Neural Comput.*, 4(2):196–210.

Azevedo, F., Carvalho, L., Grinberg, L., Farfel, J., Ferretti, R., Leite, R., Filho, W., Lent, R., and Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.*, 513(5):532–541.

Ballard, D. and Jehee, J. (2012). Dynamic coding of signed quantities in cortical feedback circuits. *Front. Psych.*, 3:254.

Banitt, Y., Martin, K., and Segev, I. (2007). A biologically realistic model of contrast invariant orientation tuning by thalamocortical synaptic depression. *J. Neurosci.*, 27(38):10230–10239.

Barlow, H. (1989). Unsupervised learning. *Neural Comput.*, 1(3):295–311.

Bartfeld, E. and Grinvald, A. (1992). Relationships between orientation-preference pinwheels, cytochrome oxidase blobs, and ocular-dominance columns in primate striate cortex. *Proc. Natl. Acad. Sci. USA*, 89(24):11905–11909.

Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., and Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Batschelet, E. (1981). *Circular statistics in biology*. London: Academic.

Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vision*, 5(6):9–9.

Bienenstock, E., Cooper, L., and Munro, P. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2(1):32–48.

Bonin, V., Histed, M., Yurgenson, S., and Reid, R. (2011). Local diversity and fine-scale organization of receptive fields in mouse visual cortex. *J. Neurosci.*, 31(50):18506–18521.

Borghuis, B., Ratliff, C., Smith, R., Sterling, P., and Balasubramanian, V. (2008). Design of a neuronal array. *J. Neurosci.*, 28(12):3178–3189.

Burkitt, A. (2006a). A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biol. Cybern.*, 95(1):1–19.

Burkitt, A. (2006b). A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties. *Biol. Cybern.*, 95(2):97–112.

Cadieu, C., Hong, H., Yamins, D., Pinto, N., Ardila, D., Solomon, E., Majaj, N., and DiCarlo, J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.*, 10(12):e1003963.

Callaway, E. (2001). Neural mechanisms for the generation of visual complex cells. *Neuron*, 32(3):378–380.

Carandini, M. (2006). What simple and complex cells compute. *J. Physiol.*, 577(2):463–466.

Carandini, M., Demb, J., Mante, V., Tolhurst, D., Dan, Y., Olshausen, B., Gallant, J., and Rust, N. (2005). Do we know what the early visual system does? *J. Neurosci.*, 25(46):10577–10597.

Carandini, M. and Heeger, D. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.*, 13(1):51.

Chapman, B. and Stryker, M. (1993). Development of orientation selectivity in ferret visual cortex and effects of deprivation. *J. Neurosci.*, 13(12):5251–5262.

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

Cloherty, S. and Ibbotson, M. (2014). Contrast-dependent phase sensitivity in V1 but not V2 of macaque visual cortex. *J. Neurophysiol.*, 113(2):434–444.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314.

Cooper, L. and Bear, M. (2012). The BCM theory of synapse modification at 30: interaction of theory with experiment. *Nat. Rev. Neurosci.*, 13(11):798.

Dayan, P. and Abbott, L. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT Press Cambridge, MA.

De Valois, R., Albrecht, D., and Thorell, L. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Res.*, 22(5):545–559.

Einhäuser, W., Kayser, C., König, P., and Körding, K. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur. J. Neurosci.*, 15(3):475–486.

Feldmeyer, D., Lübke, J., Silver, R., and Sakmann, B. (2002). Synaptic connections between layer 4 spiny neurone-layer 2/3 pyramidal cell pairs in juvenile rat barrel cortex: physiology and anatomy of interlaminar signalling within a cortical column. *J. Physiol.*, 538(3):803–822.

Feldmeyer, D., Roth, A., and Sakmann, B. (2005). Monosynaptic connections between pairs of spiny stellate cells in layer 4 and pyramidal cells in layer 5A indicate that lemniscal and paralemniscal afferent pathways converge in the infragranular somatosensory cortex. *J. Neurosci.*, 25(13):3423–3431.

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

Ferster, D. (1988). Spatially opponent excitation and inhibition in simple cells of the cat visual cortex. *J. Neurosci.*, 8(4):1172–1180.

Ferster, D., Chung, S., and Wheat, H. (1996). Orientation selectivity of thalamic input to simple cells of cat visual cortex. *Nature*, 380(6571):249–252.

Field, G. and Chichilnisky, E. (2007). Information processing in the primate retina: circuitry and coding. *Ann. Rev. Neurosci.*, 30:1–30.

Finn, I., Priebe, N., and Ferster, D. (2007). The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. *Neuron*, 54(1):137–152.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.*, 3(2):194–200.

Franciosini, A., Boutin, V., and Perrinet, L. (2019). Modelling complex cells of early visual cortex using predictive coding. In *28th Annual Computational Neuroscience Meeting*. `https://laurentperrinet.github.io/publication/franciosini-perrinet-19-cns/franciosini-perrinet-19-cns.pdf`.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 36(4):193–202.

Hawken, M., Parker, A., and Lund, J. (1988). Laminar organization and contrast sensitivity of direction-selective cells in the striate cortex of the old world monkey. *J. Neurosci.*, 8(10):3541–3548.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Visual Neurosci.*, 9(2):181–197.

Hinton, G., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554.

Hirsch, J., Alonso, J., Reid, R., and Martinez, L. (1998). Synaptic integration in striate cortical simple cells. *J. Neurosci.*, 18(22):9517–9528.

Hirsch, J., Martinez, L., Pillai, C., Alonso, J., Wang, Q., and Sommer, F. (2003). Functionally distinct inhibitory neurons at the first stage of visual cortical processing. *Nat. Neurosci.*, 6(12):1300.

Hodgkin, A. and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4):500–544.

Hoffmann, K. and Stone, J. (1971). Conduction velocity of afferents to cat visual cortex: a correlation with cortical receptive field properties. *Brain Res.*, 32(2):460–466.

Hosoya, H. and Hyvärinen, A. (2016). Learning visual spatial pooling by strong pca dimension reduction. *Neural Comput.*, 28(7):1249–1264.

Hubel, D. (1995). *Eye, brain, and vision.* Scientific American Library/Scientific American Books.

Hubel, D. and Wiesel, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, 148(3):574–591.

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160(1):106–154.

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, 195(1):215–243.

Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.*, 12(7):1705–1720.

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

Jehee, J. and Ballard, D. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Comput. Biol.*, 5(5):e1000373.

Jehee, J., Rothkopf, C., Beck, J., and Ballard, D. (2006). Learning receptive fields using predictive feedback. *J. Physiol.-Paris*, 100(1):125–132.

Jin, J., Wang, Y., Swadlow, H., and Alonso, J. (2011). Population receptive fields of on and off thalamic inputs to an orientation column in visual cortex. *Nat. Neurosci.*, 14(2):232–240.

Jin, J., Weng, C., Yeh, C., Gordon, J., Ruthazer, E., Stryker, M., Swadlow, H., and Alonso, J. (2008). On and off domains of geniculate afferents in cat primary visual cortex. *Nat. Neurosci.*, 11(1):88–94.

Jones, J. and Palmer, L. (1987a). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1233–1258.

Jones, J. and Palmer, L. (1987b). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1187–1211.

Kandel, E., Schwartz, J., Jessell, T., Siegelbaum, S., and Hudspeth, A. (2013). *Principles of Neural Science, Fifth Edition*. McGraw-Hill Education, New York.

Kremkow, J., Perrinet, L., Monier, C., Alonso, J., Aertsen, A., Frégnac, Y., and Masson, G. (2016). Push-pull receptive field organization and synaptic depression: mechanisms for reliably encoding naturalistic stimuli in V1. *Front. Neural Circuits*, 10:37.

Kretz, R., Rager, G., and Norton, T. (1986). Laminar organization of ON and OFF regions and ocular dominance in the striate cortex of the tree shrew (tupaia belangeri). *J. Comp. Neurol.*, 251(1):135–145.

Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A., and Wiskott, L. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Trans. Pattern. Anal. Mach. Intell.*, 35(8):1847–1871.

Law, C. and Cooper, L. (1994). Formation of receptive fields in realistic visual environments according to the bienenstock, cooper, and munro (BCM) theory. *Proc. Natl. Acad. Sci. USA*, 91(16):7797–7801.

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10):1995.

Lee, W., Bonin, V., Reed, M., Graham, B., Hood, G., Glattfelder, K., and Reid, R. (2016). Anatomy and function of an excitatory network in the visual cortex. *Nature*, 532(7599):370–374.

Levy, W. and Baxter, R. (1996). Energy efficient neural codes. *Neural Comput.*, 8(3):531–543.

Lian, Y., Meffin, H., Grayden, D., Kameneva, T., and Burkitt, A. (2019). Towards a biologically plausible model of LGN-V1 pathways based on efficient coding. *Front. Neural Circuits*, 13:13.

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proc. Natl. Acad. Sci.*, 83(19):7508–7512.

Lund, J. (1984). Spiny stellate neurons. *Cereb. Cortex*, 1:255–308.

Ma, L. and Zhang, L. (2007). A hierarchical generative model for overcomplete topographic representations in natural images. In *International Joint Conference on Neural Networks*, pages 1198–1203. IEEE.

Mardia, K. (1972). *Statistics of directional data*. London: Academic.

Martin, K. (2002). Microcircuits in visual cortex. *Curr. Opin. Neurobio.*, 12(4):418–425.

Martinez, L. and Alonso, J. (2001). Construction of complex receptive fields in cat primary visual cortex. *Neuron*, 32(3):515–525.

Martinez, L. and Alonso, J. (2003). Complex receptive fields in primary visual cortex. *Neuroscientist*, 9(5):317–331.

Martinez, L., Wang, Q., Reid, R., Pillai, C., Alonso, J., Sommer, F., and Hirsch, J. (2005). Receptive field structure varies with layer in the primary visual cortex. *Nat. Neurosci.*, 8(3):372–379.

McFarland, J., Cui, Y., and Butts, D. (2013). Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Comput. Biol.*, 9(7):e1003143.

Meffin, H., Hietanen, M., Cloherty, S., and Ibbotson, M. (2015). Spatial phase sensitivity of complex cells in primary visual cortex depends on stimulus contrast. *J. Neurophysiol.*, 114(6):3326–3338.

Mercer, A., West, D., Morris, O., Kirchhecker, S., Kerkhoff, J., and Thomson, A. (2005). Excitatory connections made by presynaptic cortico-cortical pyramidal cells in layer 6 of the neocortex. *Cereb. Cortex*, 15(10):1485–1496.

Mountcastle, V., Berman, A., and Davies, P. (1955). Topographic organization and modality representation in first somatic area of cat's cerebral cortex by method of single unit analysis. *Am. J. Physiol.*, 183(464):10.

Movshon, J., Thompson, I., and Tolhurst, D. (1978a). Receptive field organization of complex cells in the cat's striate cortex. *J. Physiol.*, 283(1):79–99.

Movshon, J., Thompson, I., and Tolhurst, D. (1978b). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol.*, 283(1):53–77.

Muly, E. and Fitzpatrick, D. (1992). The morphological basis for binocular and ON/OFF convergence in tree shrew striate cortex. *J. Neurosci.*, 12(4):1319–1334.

Ohki, K., Chung, S., Ch'ng, Y., Kara, P., and Reid, R. (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. (2006). Highly ordered arrangement of single neurons in orientation pinwheels. *Nature*, 442(7105):925.

Olshausen, B., Cadieu, C., and Warland, D. (2009). Learning real and complex overcomplete representations from the statistics of natural images. In *SPIE Proceedings*, volume 7446: Wavelets XIII, (V.K. Goyal, M. Papadakis, D. van de Ville, Eds.), page 74460S.

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

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

Peters, A. and Payne, B. (1993). Numerical relationships between geniculocortical afferents and pyramidal cell modules in cat primary visual cortex. *Cereb. Cortex*, 3(1):69–78.

Potjans, T. and Diesmann, M. (2014). The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cereb. Cortex*, 24(3):785–806.

Priebe, N. (2016). Mechanisms of orientation selectivity in the primary visual cortex. *Ann. Rev. Vis. Sci.*, 2:85–107.

Ratliff, C., Borghuis, B., Kao, Y., Sterling, P., and Balasubramanian, V. (2010). Retina is structured to process an excess of darkness in natural scenes. *Proc. Natl. Acad. Sci. USA*, 107(40):17368–17373.

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

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

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

Rosenbaltt, F. (1957). The perceptron - a perciving and recognizing automation. Technical report, Report 85-460-1 Cornell Aeronautical Laboratory, Ithaca.

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

Rubin, D., Van Hooser, S., and Miller, K. (2015). The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Schiller, P., Finlay, B., and Volman, S. (1976a). Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields. *J. Neurophysiol.*, 39(6):1288–1319.

Schiller, P., Finlay, B., and Volman, S. (1976b). Quantitative studies of single-cell properties in monkey striate cortex. II. Orientation specificity and ocular dominance. *J. Neurophysiol.*, 39(6):1320–1333.

Schwartz, O., Pillow, J., Rust, N., and Simoncelli, E. (2006). Spike-triggered neural characterization. *J. Vision*, 6(4):484–507.

Sclar, G. and Freeman, R. (1982). Orientation selectivity in the cat's striate cortex is invariant with stimulus contrast. *Exp. Brain Res.*, 46(3):457–461.

Sherman, S. and Guillery, R. (1996). Functional organization of thalamocortical relays. *J. Neurophysiol.*, 76(3):1367–1395.

Silberberg, G., Grillner, S., LeBeau, F., Maex, R., and Markram, H. (2005). Synaptic pathways in neural microcircuits. *Trends Neurosci.*, 28(10):541–551.

Skottun, B., Bradley, A., Sclar, G., Ohzawa, I., and Freeman, R. (1987). The effects of contrast on visual orientation and spatial frequency discrimination: a comparison of single cells and behavior. *J. Neurophysiol.*, 57(3):773–786.

Skottun, B., De Valois, R., Grosof, D., Movshon, J., Albrecht, D., and Bonds, A. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Res.*, 31(7-8):1078–1086.

Spratling, M. (2008). Reconciling predictive coding and biased competition models of cortical function. *Front. Comput. Neurosci.*, 2:4.

Strata, P. and Harvey, R. (1999). Dale's principle. *Brain Res. Bull.*, 50(5):349–350.

Swadlow, H. (1983). Efferent systems of primary visual cortex: a review of structure and function. *Brain Res. Rev.*, 6(1):1–24.

Tadmor, Y. and Tolhurst, D. (2000). Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Res.*, 40(22):3145–3157.

Tang, J., Jimenez, S., Chakraborty, S., and Schultz, S. (2016). Visual receptive field properties of neurons in the mouse lateral geniculate nucleus. *PLoS ONE*, 11(1):e0146017.

Tarczy-Hornoch, K., Martin, K., Stratford, K., and Jack, J. (1999). Intracortical excitation of spiny neurons in layer 4 of cat striate cortex in vitro. *Cereb. Cortex*, 9(8):833–843.

Thomson, A. and Bannister, A. (1998). Postsynaptic pyramidal target selection by descending layer III pyramidal axons: dual intracellular recordings and biocytin filling in slices of rat neocortex. *J. Neurosci.*, 84(3):669–683.

Thomson, A. and Bannister, A. (2003). Interlaminar connections in the neocortex. *Cereb. Cortex*, 13(1):5–14.

Troy, J., Oh, J., and Enroth-Cugell, C. (1993). Effect of ambient illumination on the spatial properties of the center and surround of Y-cell receptive fields. *Visual Neurosic.*, 10(4):753–764.

Turrigiano, G. (2011). Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Ann. Rev. Neurosci.*, 34:89–103.

Van Hateren, J. and Van Der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biol. Sci*, 265(1394):359–366.

Van Kleef, J., Cloherty, S., and Ibbotson, M. (2010). Complex cell receptive fields: evidence for a hierarchical mechanism. *J. Physiol.*, 588(18):3457–3470.

Wang, W., Jones, H., Andolina, I., Salt, T., and Sillito, A. (2006). Functional alignment of feedback effects from visual cortex to thalamus. *Nat. Neurosci.*, 9(10):1330–1336.

Willmore, B., Bulstrode, H., and Tolhurst, D. (2012). Contrast normalization contributes to a biologically-plausible model of receptive-field development in primary visual cortex (V1). *Vision Res.*, 54:49–60.

Wiltschut, J. and Hamker, F. (2009). Efficient coding correlates with spatial frequency tuning in a model of V1 receptive field organization. *Visual Neurosci.*, 26(1):21–34.

Yunzab, M., Choi, V., Meffin, H., Cloherty, S., Priebe, N., and Ibbotson, M. (2019). Synaptic basis for contrast-dependent shifts in functional identity in mouse V1. *eNeuro*, 6(2).

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

Zylberberg, J., Murphy, J., and DeWeese, M. (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 Comput. Biol.*, 7(10):e1002250.