Title
From scenes to spikes: understanding vision from the outside in

Author
Zylberberg, Joel Leon

Publication Date
2012

Peer reviewed|Thesis/dissertation

From scenes to spikes: understanding vision from the outside in

by

Joel Leon Zylberberg

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Physics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Assistant Professor Michael DeWeese, Chair
Professor Bruno Olshausen
Professor Saul Perlmutter

Fall 2012

From scenes to spikes

by

Joel Leon Zylberberg

Abstract

From scenes to spikes
by
Joel Leon Zylberberg
Doctor of Philosophy in Physics
University of California, Berkeley
Assistant Professor Michael DeWeese, Chair

The human genome (containing $\sim 10^{10}$ bits of information) is unlikely to fully specify the connectivity between neurons in our brains – such a "wiring diagram" requires $\sim 10^{14}$ bits. Physiological evidence suggests that the genome instead specifies plasticity rules through which the brain self-organizes in response to experience. As systems neuroscientists, we seek to understand those rules and, by extension, our brains. In this thesis, I will use this approach to study the primary visual cortex (V1) – the brain region that receives visual inputs from the eyes, via a relay station called the lateral geniculate nucleus. I first study the statistical structure of natural images, which provide the visual experience that shapes V1. Then, I introduce a biophysically motivated model for visual cortex, which adapts to natural image statistics in order to efficiently encode them – in this case, the neural plasticity rules can be shown to optimize this "efficient" representation. I then demonstrate that this model can account for several features of V1 physiology, including the features to which V1 neurons respond ("receptive fields"), and the developmental trends in the sparseness of V1 activity. I will conclude that efficient coding models can be implemented within the constraints imposed by the neural substrate, and that efficient coding principles may yield a parsimonious systems-level understanding of visual cortex.

Acknowledgements

My scientific career has taken a somewhat meandering path. At times, it was not apparent to me, or anyone else for that matter, where I would end up, or when, or why. As such, I am very grateful to the people in my life who provided patient support while I wandered into my eventual (and very happy!) current field of study. Of particular note are my fabulous partner Heather Stewart, my parents Philip and Darlene, and all of the (many!) research mentors that have invested their time in my apprenticeship only to see me leave their fields of study.

While completing this work, I have benefitted enormously from the thoughtful support of other scientists. In particular, I am grateful to Mike DeWeese and Bruno Olshausen for their roles in initiating the projects described herein, and to Fritz Sommer, Jascha Sohl-Dickstein, and the rest of the Redwood Center for Theoretical Neuroscience for helpful discussions.

There are several people who contributed technical work to this thesis. They are acknowledged in the appropriate chapters.

Thanks to Mark Paskin for creating and sharing the "Thesis in a box" LaTex template with which this document was prepared.

Finally, I would not have been able to do any of this work without the financial support I received, in the form of fellowships from the Natural Sciences and Engineering Council of Canada (NSERC), Fulbright scholarship board, and the Howard Hughes Medical Institute (HHMI). I am very grateful to these funders for the opportunities they have given me.

This thesis is dedicated to those who wander but are not lost

Contents

1 Introduction **1**

1.1 Why study vision? 1

1.2 Gross anatomy of the mammalian visual system 2

1.3 What does a neuron look like and what does it do? 4

1.4 Visual systems are (at least partially) a product of their environment 10

1.5 Outline of this thesis 11

2 Low-level statistics of natural images **13**

2.1 Natural images are surprising statistically homogeneous 13

2.2 "Dead leaf" models of the 2-point statistics of natural images 14

2.3 How does occlusion affect the 2-point statistics of natural images? . . 15

3 Using the 2-point functions of natural images to understand the peripheral visual system **23**

3.1 Redundancy is information theoretically inefficient 23

3.2 Whitening in the spatial domain: a model for RGC filtering? 25

3.3 Lateral geniculate nucleus outputs are white in the time domain . . . 26

4 Characterizing the higher-order statistics of natural images **28**

4.1 Independent Component Analysis 28

4.2 Sparse Coding 30
4.3 Might primary visual cortex use a sparse code for natural images? . . 34
4.4 Summary 38

5 (How) can biophysical neurons *learn* a sparse code for natural images? 40

5.1 Sparseness and decorrelation allow biophysical neurons learn a linear generative code for natural images 40
5.2 The Sparse And Independent Local network (SAILnet) 43
5.3 When trained on natural images, SAILnet model neurons learn the full diversity of receptive fields displayed by V1 simple cells 50
5.4 SAILnet units can exhibit a broad distribution of mean firing rates in response to natural images 52
5.5 Pairs of SAILnet units have small firing rate correlations. 55
5.6 Connectivity learned by SAILnet allows for further experimental tests of the model 56
5.7 Discussion 59

6 Why does sparseness in ferret V1 *decrease* during development? 62

6.1 Decreasing sparseness during development is in conflict with the canonical sparse coding models 62
6.2 SAILnet single- and multi-unit activity can become less sparse during receptive field formation, in agreement with V1 development 65
6.3 For some initial conditions, SAILnet single-unit activity becomes more sparse during receptive field formation 68
6.4 The diversity of receptive field shapes depends on the homeostatic set point of unit activity in SAILnet 69
6.5 Discussion 69

6.6 Methods 72

7 Conclusions **73**

7.1 Contributions of this thesis 73

7.2 Directions for future research 74

Bibliography **79**

Chapter 1

Introduction

1.1 Why study vision?

Vision is (arguably) our dominant sensory modality. As such, we are strongly motivated to understand the transformation from raw visual input, into perception of our surroundings. A detailed understanding of this process will help us to better understand our own experiences, and to make machines that can mimic our impressive visual abilities, like object recognition.

Furthermore, if we can understand our visual system mechanistically, we will be better able to diagnose and treat visual dysfunctions, as the mapping from symptom to disease will be more apparent. As an added medical bonus, a detailed understanding of how visual stimuli are encoded by neurons may help to create prosthetic technologies that can interface with the brain. For example, consider a patient with damaged eyes, but an otherwise fully functioning visual system. Hypothetically, one could replace their eyes with cameras, but how would those cameras communicate with the brain?

Somewhat more pedantically, visual systems provide a very appealing preparation for understanding neural systems more generally: for those seeking to understand the input-output relationships in neural systems, and thus the computations that they perform, the visual system, having a well-defined and controllable input space, is a natural starting point. In other words, with vision, we can control the input by controlling the stimulus shown to the animal, and study the corresponding neural responses. So long as there are universal aspects to the computations performed by neural circuits, the results from visual-neural studies should inform our understanding of other sensory and neural systems.

Thusly motivated, I will begin this introductory chapter by summarizing the gross anatomy of the mammalian visual system, the physiological properties of visual cortical neurons, and the characterization thereof. I will then summarize evidence which

suggests that the visual cortical physiology is a function of the visual stimuli experienced by the animal during development. This will motivate the technical work presented in the later chapters of this thesis, in which I will attempt to understand the origins of the statistical properties of natural images, and how the constraints implicit in the architecture of the visual cortex affect its ability to adapt to natural scene statistics. This work, and the literature reviewed herein, provide some insight into both *how* certain aspects of the visual physiology arise, and *why* those properties are useful to the animal.

1.2 Gross anatomy of the mammalian visual system

In order to understand the input-output properties of the primary visual cortex, we must first understand where those inputs come from. For our purposes, a somewhat cursory knowledge of the mammalian visual anatomy will suffice, such as that of [Dayan and Abbott, 2001] (Fig. 1.1). In very gross terms, people have two eyes, into which light from the surrounding environment enters, via the pupil, and forms an image on the retina, the photosensitive back surface of the eyeball. Intriguingly, the retina is "inside out", with the photoreceptors lying on the back surface. This means that light must travel through the full thickness of the retina before it can be detected by the photoreceptor (rod and cone) cells.

The photoreceptor cells encode the rate of photon capture by varying (along a continuum) their membrane potential, resulting in graded variation of the amount of glutamate (a neurotransmitter) they release into synaptic junctions between themselves and the bipolar cells, which then project to the retinal ganglions cells (RGCs), whose axonal projections (the "output" side of the cell) form the optic nerve. The bipolar cell physiology, and the lateral connectivity within the retina (by, for example, the horizontal, and amacrine, cells) are not particularly germane to this work. For an excellent overview of retinal anatomy and physiology, interested readers should consult WEBVISION [Kolb *et al.*, 2011].

The graded response of the photoreceptor cells, wherein the output varies along a continuum is in distinct contrast to the behavior of the majority of the neurons in the central nervous system, which output only punctate "spikes" of electrical activity called action potentials. Spiking cells include the RGCs, and all cortical neurons, and we will discuss them in more detail below.

The optic nerve projects to the lateral geniculate nucleus (LGN), which then projects to the primary visual cortex (V1).

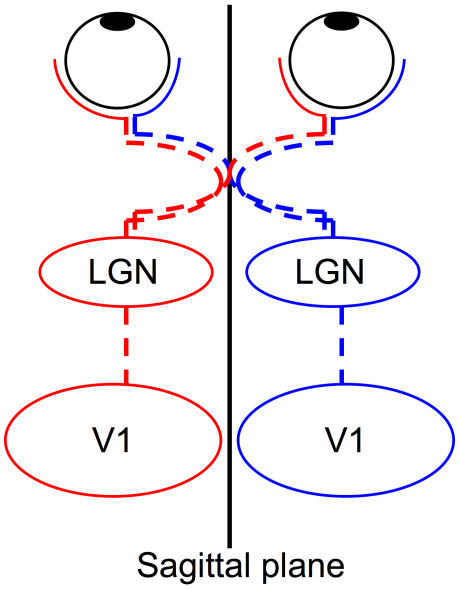

Figure 1.1: **Gross anatomy of the peripheral mammalian visual system.** Axons from the retinal ganglions cells form the optic nerve, which projects back into the head: neuronal projections are indicated by dashed lines in the figure. The temporal RGCs (those further from the sagittal plane) project to the ipsilateral LGN, while the nasal RGCs project to the contralateral LGN. The contralateral projections cross the midline at the optic chiasm (not labeled). These are color coded in the diagram, with the projections to the left half labeled in red and those projecting to the right half labeled in blue. The lateral geniculate nuclei project to the ipsilateral primary visual cortices such that the visual information represented in the left V1 corresponds to the right half of the visual field of view, and vice versa. This can be seen by ray-tracing from a point source in, say, the right half of the field-of-view, and noting that it is imaged onto the left halves of both retinae, while the left halves of both retinae project to the left LGN and from there to the left V1. For reference, the sagittal plane is shown: it runs vertically, between the two eyes. This figure is an adaptation of one from Dayan & Abbott.

1.3 What does a neuron look like and what does it do?

1.3.1 Neuronal anatomy and physiology physiology

To build a realistic model of how neural activities encode the stimulus, we require a basic knowledge of neural physiology: what is the "output" of a neuron, what are its "inputs", and how does it generate the output from those inputs? In this discussion, I will emphasize cortical neurons, which I will later be modeling. The same statements, however, also apply to retinal ganglion cells, and to neurons in the lateral geniculate nucleus.

While neurons come in very different shapes and sizes, cortical neurons have three main anatomical features: the axon, dendrites, and the soma [Dayan and Abbott, 2001] (Fig. 1.2).

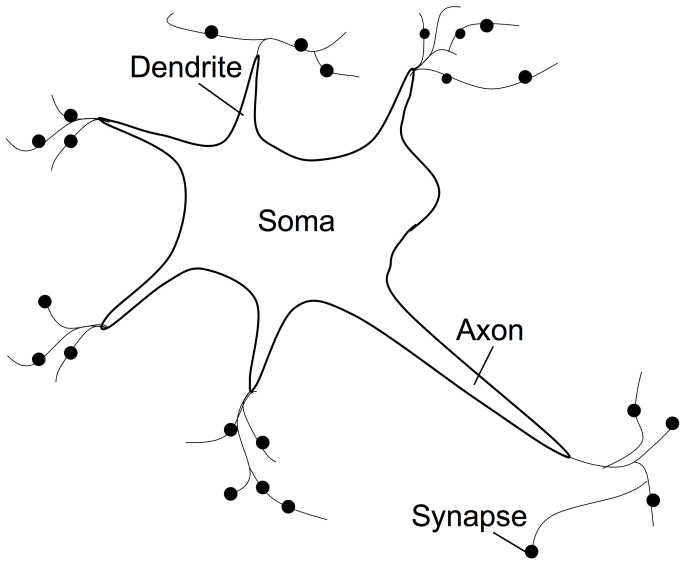

Figure 1.2: **Cartoon of a neuron**. Neurons have three major components: the axon, soma (cell body), and dendrites. The synapses (shown as black circles) are typically the connections between the axon of one cell and the dendrite of another, where the axon is the cell's "output", and the dendrites are the cell's "inputs". Both the axon and the dendrites can be highly branched, as cortical neurons are highly interconnected. This diagram is not to scale, nor does it reflect the specific morphology of any particular neuron type.

The cells have ion pumps embedded in their membranes that maintain a voltage

difference across the cell membrane: the cytoplasm of a neuron at rest is roughly $-70\ mV$ relative to the extracellular medium [Dayan and Abbott, 2001].

Inputs to the neuron (from other neurons) come into the cell at the dendrites, which form branching tree-like structures. These inputs change the membrane potential of the neuron: they can raise (excitatory inputs), or lower (inhibitory inputs) it. When the membrane potential is increased to a sufficiently high level, the neuron emits an *action potential*, or stereotyped "spike" of electrical activity that travels down the axon of the cell, where it impinges on the dendrites of neighboring neurons at connections called *synapses*. An excellent example of the time course of the membrane potential during the spiking events is given by [Mainen and Sejnowski, 1995].

The strengths of the synaptic connections between neurons change in response to the relationships between the pre- and post-synaptic activity; this is believed to be one of the major ways that learning happens in the brain [Dayan and Abbott, 2001; Dan and Poo, 2006; Feldman, 2009].

The soma, or body of the cell, lies between the dendrites and the axon, and is thought to play a major role in the computations performed by the cell, namely in determining when to spike.

The action potentials are essentially indistinguishable: while the membrane potential of the neuron is a continuous-valued function of time, all of the information output by a cortical neuron (to other neurons) is carried in the binary time-sequence of its spiking activity ("on" or "off" at each moment in time) [Dayan and Abbott, 2001]. There is much debate in the field about whether this information is carried in the *rate* at which the spikes are transmitted (rate coding) [London *et al.*, 2010; Shadlen and Newsome, 1994], or the specific times at which the spikes are emitted (spike code) [Bialek *et al.*, 1991]. Of course, it is also possible that different parts of the brain employ different coding schemes.

1.3.2 Simplified neuronal models

To formalize our study of neural systems, we require a mathematical model of the neuronal electrophysiology. The most realistic neuronal model is described by the Hodgkin-Huxley equations, which describe the behavior of several different types of voltage-gated ion channels, and their effects on the membrane potential [Izhikevich, 2007; Hodgkin and Huxley, 1952]. The Hodgkin-Huxley equations are, unfortunately, quite complicated, difficult to analyze, and computationally expensive to simulate.

To simplify our lives, and allow for efficient simulations, theorists frequently use models that dramatically simplify the neuronal physiology. For the work presented herein, I will use the leaky-integrate-and-fire model (LIF, described below) [Dayan and Abbott, 2001]. Many other neuron models exist, including quadratic integrate-

and-fire, and exponential integrate-and-fire, which are more realistic than the LIF model [Izhikevich, 2007], at the expense of added computational cost to simulate.

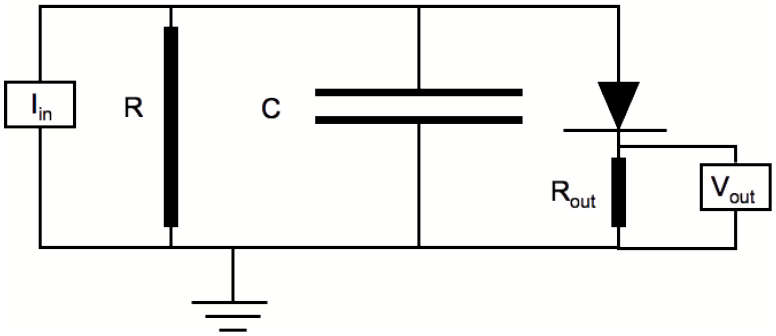

Figure 1.3: **Circuit diagram of the leaky integrate-and-fire neuron model.** The upper portion of the image (above the resistor and capacitor) represents the cytoplasm (inside) of the cell, while the lower portion (ground) represents the extracellular medium. Input currents (from action potentials impinging on dendritic synapses) cause the voltage across the capacitor to increase, while leakage (through passive ion channels in the membrane) causes the voltage to decrease. When the voltage exceeds the bias voltage of the diode, the capacitor discharges through a very small resistance $R_{out} << R$, resulting in the capacitor being grounded, and an output voltage V_{out} being generated. The diode corresponds roughly to the behavior of voltage-gated ion channels in the cell membrane of a neuron. The resistors are denoted as cylinders to highlight the fact that they correspond to transmembrane ion channels.

In a LIF neuron model (Fig. 1.3), each neuron has a "hidden variable" u associated with it, that corresponds roughly to its membrane potential (the voltage difference between the cytoplasm and extracellular medium: considering the cell membrane as a capacitor, this is proportional to the amount of charge stored in the cell). The inputs x_i that reach the cell at synapses (with weights w_i) collectively act as a current source, increasing the membrane potential, or the charge on the capacitor.

The hidden variable evolves in time via

$$\frac{du}{dt} = -u(t) + \sum_i w_i x_i(t). \tag{1.1}$$

Note that this is identical to the differential equation for the voltage across the capacitor in a parallel RC circuit with unit resistance and capacitance, and input current $\sum_i w_i x_i(t)$.

When the hidden variable exceeds some threshold θ, the neuron emits a spike of

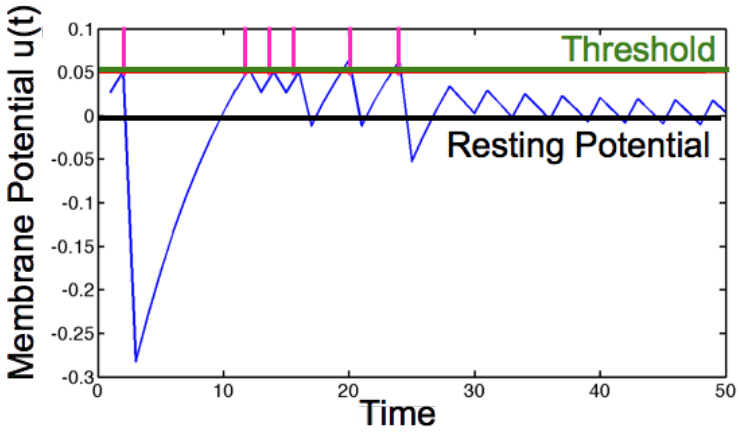

Figure 1.4: **Leaky integrate-and-fire dynamics.** The membrane potential of a LIF neuron is shown as a function of time. The neuron is receiving a constant excitatory input, with occasional pulses of inhibition that depolarize the membrane. Whenever the membrane potential exceeds the threshold, the neuron produces a spike before returning to its resting voltage (0). The spikes (shown in pink) are the only aspect of the neuronal state that are communicated to other neurons in this class of model. In the absence of inhibitory pulses, this neuron, by virtue of the constant excitatory input, would spike at regular intervals. In the (discrete time) simulation, the membrane potential was recorded after the update in each time step, so the reset to 0 after each spike is not visible. The alternative strategy, recording the membrane potential before the update, means that the spiking events, where the potential reaches threshold, would not be seen.

activity (Fig. 1.4).

$$v_{out}(t) = \Theta\left(u(t) - \theta\right), \tag{1.2}$$

where $\Theta(.)$ is the Heaviside step function. Roughly speaking, this step corresponds to the actions of voltage-gated ion channels in the cell membrane [Izhikevich, 2007]. Once the neuron spikes, the hidden variable returns to its resting value of 0.

1.3.3 Characterizing neuronal response properties

Given a neuron in the visual system, it is very natural to ask what it is, exactly, that the neuron does. As we have seen above, the trivial response is "the neuron emits spikes whenever it is sufficiently excited". Our goal, however, is to understand the neural system, and the computations performed thereby, so such a response is not particularly useful. Instead, we wish to understand which properties in the visual stimulus cause the neuron to spike; the canonical assumption in the field is that

whatever aspect of the stimulus causes the neuron to spike is presumably the feature that the neuron represents in the mental "model" of the visual input.

The canonical way to approach this problem is to assume that the neuron is a *linear time-invariant filter* (LTI filter): the rate r at which the neuron fires is a linear function of the inputs to that neuron, and that filter function does not vary with time [Dayan and Abbott, 2001].

In the case of images formed by, say, the intensity values of discrete pixels (which loosely correspond to the outputs of the retinal ganglion cells), we can write down the filter as a vector w_i. The neuron's firing rate at time k, r_k, in response to the k^{th} input image (with pixels indexed by i) x_{ik} is

$$r_k = \sum_i x_{ik} w_i. \tag{1.3}$$

In practice, this convolution is also done over the temporal sequence of images input to the eyes [Dayan and Abbott, 2001], but we consider static inputs for simplicity. The task, then, is do determine the filter w_i that defines the inputs to which the neuron is sensitive. Imagine performing an experiment where the animal is shown a set of images $\{x_{ik}\}$ with zero ensemble mean ($\sum_k x_{ik} = 0$), and we have estimated the firing rate $r_{est,k}$ of the target cell for each of these stimuli. In order to minimize the average error between our LTI model, and our data (averaged across all the stimuli we presented), we minimize the function $E = \sum_k \left(r_{est,k} - \sum_i w_i x_{ik}\right)^2$ [Dayan and Abbott, 2001]. This yields

$$\frac{\partial E}{\partial w_n} = -2 \sum_k \left(r_{est,k} - \sum_i w_i x_{ik} \right) x_{nk} = 0 \tag{1.4}$$

$$\implies \langle r_{est,k} x_{nk} \rangle = \left\langle \sum_i w_i x_{ik} x_{nk} \right\rangle, \tag{1.5}$$

where the angular brackets $\langle . \rangle$ denote the average over the ensemble of stimuli, with indices k. Now, since we assume that the filter w_i is constant over all stimulus presentations, it comes out of the average, leaving

$$\langle r_{est,k} x_{nk} \rangle = \sum_i w_i \langle x_{ik} x_{nk} \rangle . \tag{1.6}$$

If the stimulus is white noise (the spatial pixel-pixel autocorrelation function is a delta function), then $\langle x_{ik} x_{nk} \rangle = \delta_{in} \sigma^2$, where σ^2 is the variance of the (assumed to be zero-mean) pixel values. In that case, our best estimate of the filter is simply $w_i = \langle r_{est,k} x_{ik} \rangle / \sigma^2$: the activity-triggered average of our white-noise stimulus ensemble. In

practice, one counts the spikes to estimate the rate r_{est}: thus, this becomes a *spike-triggered averge*. This estimated filter function is frequently referred to as the *receptive field* of the neuron. These receptive fields are frequently measured in physiology experiments (for example, [Ringach, 2002]). Examples of receptive fields (RFs) of neurons in macaque V1 are shown in Fig. 1.5. The RFs tend to resemble edge filters, meaning that the neuron responds to places were bright and dark regions abut one another, and Gabor wavelets (1-D sinusoid multiplied by a 2-D Gaussian).

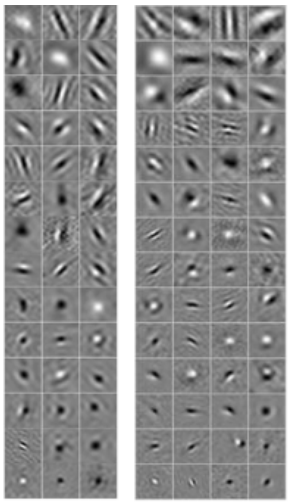

Figure 1.5: **Spike triggered average receptive fields of macaque V1 neurons in response to stimulation with natural images.** Each square in the above grid shows the receptive field of one neuron in macaque monkey primary visual cortex (courtesy of D. Ringach). The white pixels have brightness values above the mean (grey), while the black ones have brightness values below the mean. The RFs to the right of the break have an angular scale (degrees of visual angle spanned horizontally by the displayed RF window) of 0.94^o whereas those to the left of it span 1.88^o.

In the case where the pixels in the stimulus are correlated ($\langle x_{ik} x_{nk} \rangle \neq \delta_{in} \sigma^2$), the "correct" linear filter can be obtained by multiplying the STA estimate by the inverse of the autocovariance matrix of the pixel values.

Of course, given that (as we have seen above), neurons are *not* linear functions of their inputs, STA is not necessarily a good way to characterize neuronal responses, although it does yield the mean of the spike-conditional image distribution. A more sophisticated way to cast this problem is to instead ask which linear combinations of pixel values have the largest mutual information between themselves and the neuronal response (spiking or not spiking) [Sharpee *et al.*, 2004]. The weights that enter the linear combination are called the *maximally informative dimensions* (MIDs), and these have also been measured experimentally. They tend to look similar to the STA

receptive fields [Sharpee *et al.*, 2006]. Even though MID is a more sophisticated tool for characterizing neural responses, and does not rely on any shaky assumptions (like the neuron being linear or time-invariant, neither of which are strictly true [Sharpee *et al.*, 2006]), STA RFs remain the standard way to characterize the features to which neurons respond.

One of the major goals of this thesis is to extend our understanding of why these receptive fields have the shapes they have, in relation to the statistical properties of natural visual stimuli.

1.4 Visual systems are (at least partially) a product of their environment

The human genome contains $\sim 10^{10}$ bits of information: $\sim 5 \times 10^9$ base pairs [Venter *et al.*, 2001], each of which has 4 possible states, and thus contains $\log_2 4 = 2$ bits of information per base pair. At the same time, to specify which pairs of neurons are connected within the human brain would require around $\sim 10^{14}$ bits of information. There are around 10^{10} neurons in human cortex with around 10^{13} synapses connecting pairs of neurons [Shepherd, 2004]. Assuming the synapses are roughly equally distributed over the neurons, given a particular synapse on one neuron, specifying which other neuron is involved requires around $\log_2 10^{10} \approx 35$ bits per synapse.

Thus, since the genome contains less information than would be required to specify this wiring diagram (not to mention that the genome must specify other things besides just cortical wiring!), it cannot be the case that the cortical organization is strictly genetic. Instead, one might speculate that the genome contains a set of instructions by which the brain self-organizes in response to its experiences.

In the case of the visual system, then, one might reasonably expect that visual experience during the animal's maturation period would have a profound impact on the animal's visual physiology.

Indeed, this expectation is borne out by experimentation. For example, Imbert and Buisseret raised a group of kittens in the dark, thus depriving them of visual experience. They then compared the response properties of the visual cortical neurons of the dark-reared kittens to those of kittens reared in a normal, visually rich, environment. While the control kittens' visual cortical neurons were selective (as per usual) to both the orientation, and the direction of motion of a bar stimulus, the dark-reared kittens' neurons were less selective to both of these features [Imbert and Buisseret, 1975] .

Similarly, kittens that are reared in stroboscopic illumination conditions have visual cortical cells that are less selective to stimulus direction and orientation than do normally reared cats [Cynader *et al.*, 1973].

Similar results are obtained in both cats and ferrets, where orientation and direction selectivity increase during maturation [White and Fitzpatrick, 2007].

We are thus led, both by developmental studies, and by information-theoretic arguments, to the conclusion that the visual system is likely to be, at least in part, a product of the animal's environment.

At the same time, one might argue that, as a result of evolution, an animal's visual system should be organized so as to function efficiently in its environment. Indeed, this is a popular notion for understanding visual systems [Barlow, 1961; Simoncelli and Olshausen, 2001; Attneave, 1954; Linsker, 1986].

For now, it suffices to accept that the visual physiology reflects the statistics of the visual environment. We will later formulate this problem more explicitly, and discuss both previous work in this area, and the specific contributions made by this thesis work.

1.5 Outline of this thesis

The rest of this dissertation is structured as follows:

- In Ch. 2, I discuss the empirical observation that natural images seem to all have the same autocorrelation function, and theoretical work that attempts to explain this observation. At the same time, I will discuss some of my work in this area, which reveals that natural image autocorrelation functions are not affected by occlusion, the fact that objects can "hide" behind one another.

- In Ch. 3, I summarize prior work that attempts to explain the physiology of the retinal ganglion cells and the lateral geniculate nucleus in terms of the autocorrelation functions of natural images.

- Chapter 4 summarizes previous attempts to characterize the higher-order statistics of natural images, beyond the autocorrelation, or 2-point functions. I will observe in this chapter that this work may provide some deep insights into the function of the primary visual cortex.

- The theoretical models presented in Ch. 4 lack the biophysical realism needed to explain *how* the visual cortex might adapt to the statistical properties of natural images. In Ch. 5, I will discuss a resolution to this problem that, in essence, provides the first existence proof that visual cortex really could learn a sparse code for natural images, despite the locality constraints implicit in the cortical architecture.

- The existing models discussed in Ch. 4 will all predict that sparseness (to be defined later on) should increase during the adaptation process. In Ch. 6, I will confront that prediction with the empirical observation that sparseness levels in ferret visual cortex *decrease* as the animal matures. I will then provide a potential resolution to this problem, involving homeostasis, and conclude that our SAILnet model (presented in Ch. 5) might provide a parsimonious explanation for both the receptive fields of the mature V1 neurons, as well as the developmental trends leading to that state.

- Chapter 7 will summarize the main arguments in this thesis, and discuss some directions for future research that are suggested by the work presented herein.

Chapters 2, 5, and 6, are adapted from separate stand-alone manuscripts that have either been published, or submitted, at the time of the writing of this thesis. Thus, some priority claims may be inadvertently repeated in this thesis.

Chapter 2

Low-level statistics of natural images

In order to understand how our visual systems reflect the statistics of natural visual stimuli, we must first discuss those statistics. In this chapter, I will summarize both empirical, and theoretical, work on the 2-point statistics of images of natural scenes.

2.1 Natural images are surprising statistically homogeneous

Natural images are surprisingly statistically uniform. For example, the *autocorrelation function*, a measure of how much a pixel value at one location of an image can predict pixel values at other locations, is virtually universal for natural images [Ruderman and Bialek, 1994; Dong and Atick, 1995a; Field, 1987; Simoncelli and Olshausen, 2001; Torralba and Oliva, 2003] (Fig. 2.1). This is typically quantified by measuring image power spectra (Fourier transform of the autocorrelation function: recall the convolution theorem), which are well-described by scale-invariant power law functions with power $\mathcal{P}$ and spatial frequency k related by $\mathcal{P}(k) \propto k^{-\alpha}$, with exponents $\alpha \approx 2$. The exponents α vary slightly from image-to-image, and there are small differences in average exponent α between terrestrial [Field, 1987; Ruderman and Bialek, 1994] and aquatic [Balboa and Grzywacz, 2003] environments, and between natural and man-made environments [Torralba and Oliva, 2003]. Intriguingly, even radiological images like mammograms have power law power spectra [Heine and Velthuizen, 2002] (albeit with smaller α values), despite the fact that the physics of image formation are very different for radiological, and natural images. In natural images, formed by reflection of light off of surfaces, objects tend to be opaque, and thus they occlude one another, while in mammogram images, formed by the transmission of x-rays through breast tissue, objects are more transmissive, and do not completely occlude one another. Similar statements can be made about other types of radiological images.

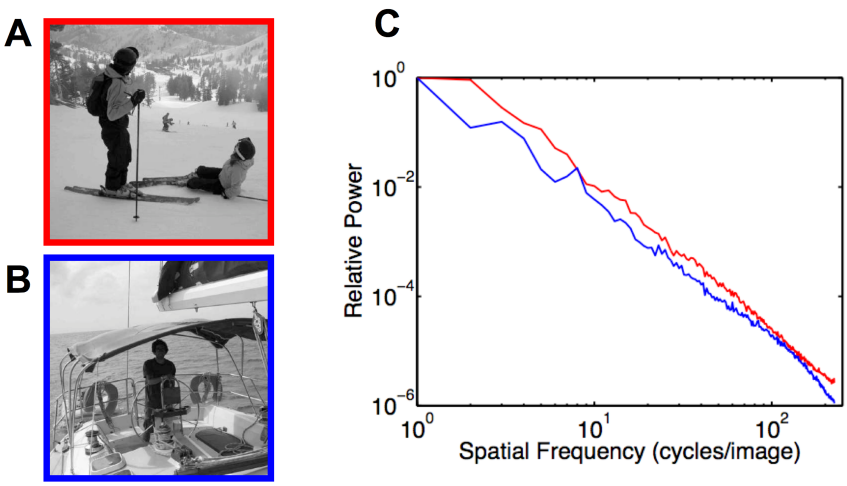

Figure 2.1: **Natural images have nearly identical power spectra.** (**A** and **B**) Two gray-scale natural images, taken by different photographers, in very different environments, have very similar rotation-averaged power spectra (**C**). The line colors in panel **C** match the borders in panels **A** and **B**.

2.2 "Dead leaf" models of the 2-point statistics of natural images

Using the intuition that the environment is composed of distinct objects, Ruderman studied a "dead leaves" model [Matheron, 1968; Bordenave *et al.*, 2006] for natural scenes, in which images are created by sequentially placing opaque, potentially overlapping circles of random brightness in random locations on a 2-dimensional image plane [Ruderman, 1997] (Fig. 2.2). Using analytical calculations he demonstrated that, so long as the diameters s of the circles follow a power law distribution with probabilities $p(s) \propto s^{-(3+\eta)}$, the images exhibit power law correlation functions, $C(q) \propto q^{-\eta}$, where q is the separation between pixels, and power law power spectra, $\mathcal{P}(k) \propto k^{-(2-\eta)}$. If the circle sizes are drawn from other distributions, Ruderman's analytic calculations suggest that the power spectra that could be made to differ from a power law, contrary to the old notion [Carlson, 1978] that the $1/k^2$ power spectra result from the mere presence of edges, each of which has a $1/k^2$ 1-dimensional power spectrum (*cf.* Balboa et al. [Balboa *et al.*, 2001]). More recently, Balboa et al. [Balboa *et al.*, 2001] simulated the examples presented by Ruderman – Ruderman [Ruderman, 1997] did not report simulation results – including images with the exponential distribution of object sizes that was claimed [Ruderman, 1997] to yield non-power-law power spectra. They found that these images had nearly power law power spectra, and subsequently re-iterated the claim that occlusion, and not object

size distributions, are the cause of power law power spectra in natural images.

This "edges vs. size distributions" debate was subsequently resolved when Hsiao and Milane demonstrated, via numerical simulations, that dead leaf models with partially transparent objects (and thus only partial occlusion) whose sizes follow a power law distribution yield power law power spectra, and that dead leaf models with opaque objects from other size distributions can have non power-law power spectra [Hsiao and Millane, 2005]. In other words, occlusion is neither necessary, nor sufficient, to yield power law image power spectra. In that same paper, Hsiao and Milane computed the power spectra of a simplified ensemble of images (simpler than the images with partially occluding leaves that they simulated), formed by summing the intensities of different randomly placed disks. The linearity of this model makes it relatively straightforward to compute the Fourier transform of the model images, and thus to estimate the power spectra.

2.3 How does occlusion affect the 2-point statistics of natural images?

Thus, to date, the 2-point statistics of dead leaf image models have been analytically calculated for both fully opaque leaves (by Ruderman), and for fully transmissive leaves (the linear model of Hsiao and Milane). What remains is to solve for the 2-point function of images with partial occlusion, which will deepen our understanding of how opacity and image statistics inter-relate along this continuum of object properties.

Thusly motivated, we studied a generalized dead leaves model, in which the leaves have variable transparency. While general feature probabilities have been solved exactly for the fully opaque dead leaves model [Pitkow and Meister, 2012], our transparent generalization requires other methods and has not previously been systematically explored. We show herein that, so long as leaf sizes follow a power-law distribution, transparency results in an overall multiplicative factor in the 2-point function but does not change its functional (power-law) form. For other size distributions, transparency does change the form of the autocorrelation function, suggesting that power-law size distributions unify the observed power spectra of natural and radiological images.

I begin by analytically computing the 2-point functions of images in the "transmissive dead leaf" environment. For image pixels values $I(\vec{x})$, the 2-point function is given by $C(\vec{x}, \vec{x}') = \langle I(\vec{x})I(\vec{x}')\rangle = C(|\vec{x} - \vec{x}'|)$, where the average denoted by the triangular brackets is taken over different pairs of image points in the ensemble of images, and the second step stems from the fact that, since the model world is invariant under both translations and rotations, the 2-point function depends only on the distance $|\vec{x} - \vec{x}'| = q$ between sample points.

The image is formed by taking a circle whose diameter s is drawn from some

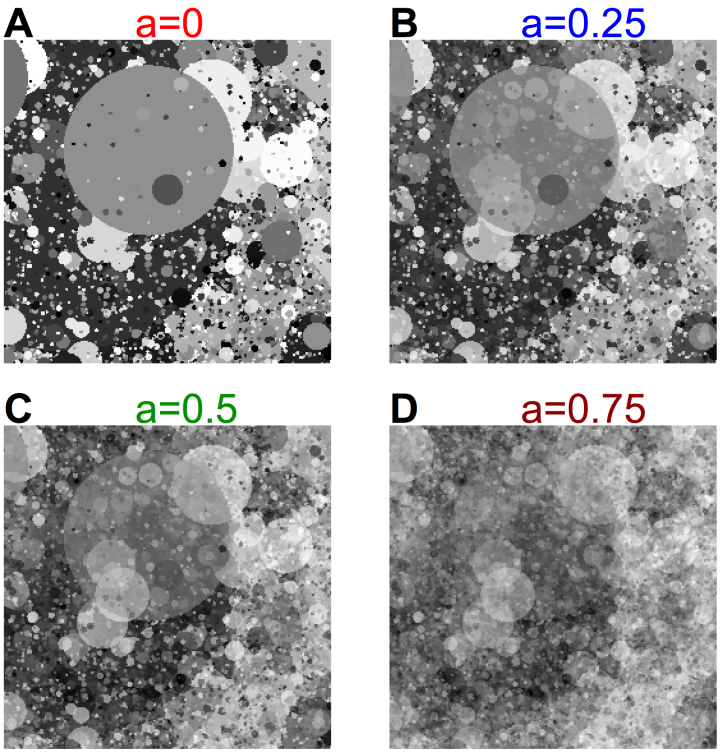

Figure 2.2: **Example opaque and transmissive dead leaf images.** Example images from the dead leaf model, with circle diameters drawn from the distribution $p(s) \propto s^{-3.2}$ for $s > s_0 = 1$ pixel, and circle brightnesses drawn from a uniform distribution over the range $[-1, 1]$. When the circles are partially transparent, but all other parameters are the same, the model images are visibly different. In the case of the transmissive ensembles (**B,C,D**), leaves behind the frontmost one are visible. The converse is true of the opaque dead leaf images (**A**), where only the frontmost surface is visible. As the leaves become more transmissive ($a \to 1$), the images become more like pink Gaussian noise, as one might expect from the central limit theorem..

distribution, with brightness value b, and transparency a, and placing it randomly on a surface of diameter L. The brightnesses b will be drawn from a zero-mean distribution, and the transparencies $a \in [0, 1]$ can also be random. A value $a = 1$ specifies a fully transparent (invisible) circle, while a value of $a = 0$ specifies a fully opaque circle, as in Ruderman's model. After the new circle is added, The resultant pixel value $I_f(\vec{x})$ at a point $\vec{x}$ that falls within the circle is

$$I_f(\vec{x}) = (1 - a)b + aI_i(\vec{x}), \tag{2.1}$$

where I_i and I_f refer to the pixel values before and after the circle is added, respectively. Those pixels not lying under the circle are unaffected by its addition. This process is continued ad infinitum to create model images, examples of which are shown in Fig. 2.2 (different panels show images created by circles of different opacity values a). It is trivial to show that the mean pixel value in these image ensembles is zero.

I will compute $\langle I(\vec{x})^2 \rangle$ and $C(q)$ recursively by first noting that adding another leaf to an image creates a new image from the transmissive dead leaf ensemble, and thus the (average) statistical properties must remain unchanged by this transformation [Ruderman, 1997].

Using Eq. 2.1, the pixel variance is computed to be

$$\begin{aligned} \left\langle I^2(\vec{x}) \right\rangle &= (1 - P_{in}) \left\langle I^2(\vec{x}) \right\rangle + P_{in} \left\langle (aI(\vec{x}) + (1-a)b)^2 \right\rangle \\ \implies \left\langle I^2(\vec{x}) \right\rangle &= \frac{\langle b^2 \rangle \langle (1-a)^2 \rangle}{1 - \langle a^2 \rangle}, \end{aligned} \tag{2.2}$$

where P_{in} is the probability that the point in question falls within the newly added circle. The quantity P_{in}, and thus the distribution of circle sizes, does not affect the pixel variance. It will however, affect the spatial properties of the image, including $C(q)$.

To compute $C(q)$, first consider how the pixel values of a pair of points, with separation q, are affected by the addition of a new leaf to the image. After adding the new leaf, either one, both, or neither, of the two sample points will lie under the leaf, resulting in three different possible modifications to the pixel values (Eq. 2.1). These outcomes occur with probabilities $P_1(q)$, $P_2(q)$, or $P_0(q)$, respectively, which I will later compute. Equating the 2-point functions both before, and after, the addition of a new leaf to the image,

$$\begin{aligned} C(q) &= P_0(q)C(q) + P_1(q) \left\langle [aI(\vec{x}) + (1-a)b] \, I(\vec{x}') \right\rangle \\ + \quad & P_2(q) \left\langle [aI(\vec{x}) + (1-a)b] \, [aI(\vec{x}') + (1-a)b] \right\rangle . \end{aligned} \tag{2.3}$$

Recalling the definition of the autocorrelation function and the normalization $P_0(q) + P_1(q) + P_2(q) = 1$, a bit of algebra yields

$$C(q) = \frac{\langle b^2 \rangle \langle (1-a)^2 \rangle \, P_2(q)}{P_1(q) \langle 1 - a \rangle + P_2(q) \langle 1 - a^2 \rangle}. \tag{2.4}$$

The quantities $\langle b^2 \rangle$, $\langle a^2 \rangle$, and $\langle a \rangle$ depend on the distributions of circle brightnesses and opacities.

I will now compute the functions $P_1(q)$ and $P_2(q)$, thus finishing the calculation of $C(q)$. In order to calculate $P_1(q)$, I begin by defining $P^\star = \langle s^2 \rangle / L^2$ to be the probability that any given point in the image falls within a newly-deposited circle. The probability $P_1(q)$ that either point, but not both, falls within the circle is thus $P_1(q) = 2\left(P^\star - P_2(q)\right)$, where the factor of 2 comes in because there are two such points to consider.

In order to determine the probability $P_2(q)$, note that, for a circle of diameter s, given that one particular point $\vec{x}$ is within the circle (which occurs with probability s^2/L^2), the probability that another point, a distance q away, is also within the circle, is given by $g(q/s \in [0,1]) = \frac{2}{\pi}\left[\cos^{-1}(q/s) - (q/s)\sqrt{1-(q/s)^2}\right]$ [Ruderman, 1997], and thus

$$P_2(q) = \int_0^\infty \frac{s^2}{L^2} g(q/s) p(s) ds. \tag{2.5}$$

For a power law size distribution $p(s) = (A/s_0)(s/s_0)^{-\alpha}$ (for power $\alpha > 3$ and normalizing constant A, and above the small-size cutoff s_0), the change of variables $u = s/q$ in the above integral yields

$$P_2(q) = A\left(\frac{s_0}{L}\right)^2 \left(\frac{q}{s_0}\right)^{-(\alpha-3)} \int_1^\infty g(1/u) u^{2-\alpha} du. \tag{2.6}$$

Define the integral to be $B(\alpha)$. For pixel separations much larger than the small-size cutoff of the leaf diameter distribution, $q \gg s_0$ (in which case $P^\star = \frac{A}{\alpha-3}\left(\frac{s_0}{L}\right)^2 \gg P_2(q)$), Eqs. 2.4 and 2.6 then yield

$$C(q) = \frac{B(\alpha)\,(\alpha-3)\,\langle b^2 \rangle\,\langle (1-a)^2 \rangle}{2\,\langle 1-a \rangle}\left(\frac{q}{s_0}\right)^{-(\alpha-3)}, \tag{2.7}$$

resulting in an image power spectrum $\mathcal{P}(k) \propto \frac{\langle b^2 \rangle \langle (1-a)^2 \rangle}{\langle 1-a \rangle} k^{-(5-\alpha)}$, in which the opacity affects the power spectrum only as a multiplicative prefactor. In the case that $a = 0$ for all circles (opaque limit), this result is equal to that of Ruderman, as it must be. Also note that, as one might expect, the 2-point function does not depend on the size L of the image surface.

To confirm my analytical calculations, I simulated 500-frame ensembles of 256×256 pixel images, using the procedure described in Eq. 2.1: circles of random size (following a power law distribution $p(s) \propto s^{-3.2}$ above the cutoff of $s_0 = 1$ pixel), brightness, and position, were iteratively placed on the image frame to build up the images. For each frame, 10^6 circles were deposited: this number was chosen to be several times larger than the number needed to cover the surface of the image with a stack of leaves several tens deep.

To avoid "edge effects", the circle centers were allowed to fall anywhere up to $256 + s/2$ away from the center of the image frame, where s is the circle diameter in pixels. Since the maximum circle size in our simulation was 10^8 pixels, this must be done carefully in order for the simulation to run in a reasonable amount of time. The maximum circle size was chosen to be so large because prior work [Lee *et al.*, 2001] shows that, when simulating dead leaf images, the functional form of the measured autocorrelation function depends on this maximum size, and approaches the analytically calculated size only in the $s_{max} \to \infty$ limit. Briefly, since the power-law distribution has such a heavy tail, it contains a non-negligible (but small) number of very large leaves, which are responsible for much of the long-range correlations in the images. Thus, one must be careful to sample the tail of this distribution in order to observe the "correct" correlation function.

I then measured the difference functions $D(q) = \left\langle |I(\vec{x}) - I(\vec{x'})|^2 \right\rangle = 2\left\langle I(\vec{x})^2 \right\rangle - 2C(q)$ for each image ensemble. The difference function is clearly related to the autocorrelation function $C(q)$, but is easier to measure as it is unaffected by the mean values of the individual images [Ruderman, 1997]. I fit the measured difference functions to power law functions of the form $D(q) = \eta \times q^{\mu} + \nu$, as is suggested by our analytical calculations (above). The best-fit parameters (η, μ, ν) for the opaque image ensemble (a=0) were $(-0.48 \pm 0.01, -0.24 \pm 0.04, 0.69 \pm 0.03)$, while the a=0.25 ensemble yielded $(\eta, \mu, \nu) = (-0.32 \pm 0.01, -0.23 \pm 0.03, 0.41 \pm 0.02)$, and the a=0.5 and a=0.75 ensembles yielded $(-0.191 \pm 0.004, -0.22 \pm 0.03, 0.23 \pm 0.01)$ and $(-0.086 \pm 0.002, -0.21 \pm 0.02, 0.098 \pm 0.005)$, respectively. These values are in reasonably good agreement with the analytical caculations (above) that suggest $\mu = -0.2$ for all image ensembles, and $\nu = \{0.67, 0.4, 0.22, 0.095\}$ for the $a = \{0, 0.25, 0.5, 0.75\}$ image ensembles, respectively.

The correlation functions displayed in Fig. 2.3 are the measured difference functions subtracted from the constants ν measured in the fit: $C(q) = [\nu - D(q)]/2$. As suggested by my analytical calculation, these correlation functions are power-law functions of distance (appearing linear on the log-log plot), and differ by a multiplicative constant. Similarly, the power spectra of the image ensembles, shown in Fig. 2.3, differ only by a multiplicative constant, and are power-law functions of spatial frequency. Above approximately 30 cycles/image, a kink can be seen in the power spectra. This occurs because the higher spatial frequencies correspond to smaller length scales, on which our $q \gg s_0$ approximation fails, and deviations from power-law behavior are thus anticipated.

To demonstrate that the leaf opacity can affect the functional form of the power spectrum, I repeated the above calculations with size distribution $p(s) = \delta(s - s^{\star})$,

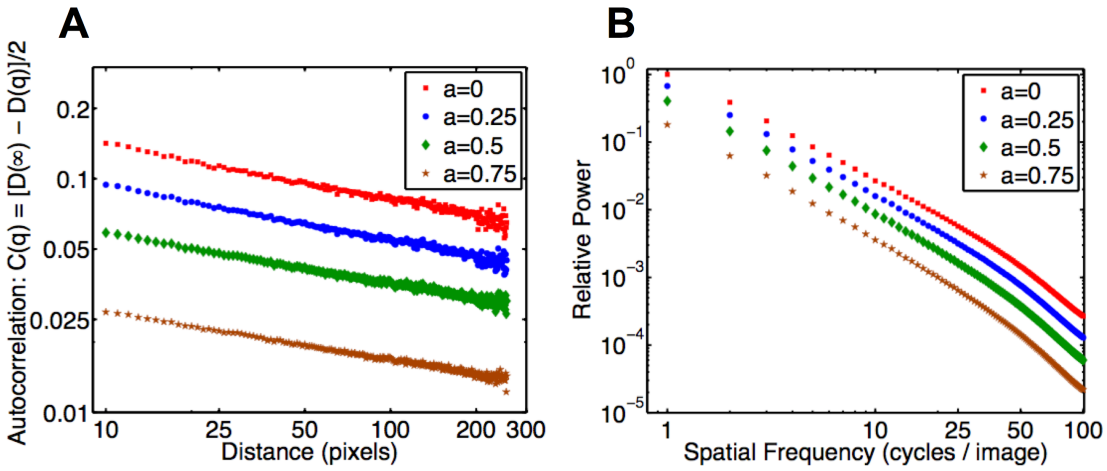

Figure 2.3: **The 2-point statistics of the dead leaf image ensembes differ only by a multiplicative constant for power-law object size distributions.** (**A**) The autocorrelation functions for the ensembles of dead leaf images shown in Fig. 2.2 are the same up to a multiplicative constant. The 2-point functions are well-described by power law functions of distance, with power ~ -0.2, in good agreement with the analytical calculation (see text for calculation and simulation analysis details). Similarly, the power spectra (**B**) of the different image ensembles are roughly power-law functions (appearing linear on a log-log plot), and are the same up to a multiplicative constant. At higher spatial frequencies, corresponding to shorter length scales, the $q \gg s_0$ approximation becomes worse, and slight deviations from power-law behavior are apparent in the power spectra.

in which case the correlation function is given by

$$C_\delta(q) = \frac{\langle b^2 \rangle \left\langle (1-a)^2 \right\rangle g(q/s^\star)}{2\langle 1-a \rangle - \left\langle (1+a)^2 \right\rangle g(q/s^\star)}, \tag{2.8}$$

which depends non-trivially on a: for $q > s^\star, g(q/s^\star) = 0$, and the correlation function vanishes, which means that the large-q limit in which Eq. 2.7 was derived is irrelevant for delta-function size distributions.

In Fig. 2.4, I demonstrate that the 2-point function is affected substantially by leaf opacity, for delta-function leaf size distributions. The procedures used to generate the data shown in Fig. 2.4 were very similar to those (discussed above) for the power-law object size distribution.

To summarize, for the special case of power-law object size distributions, object opacity does not affect the form of either the 2-point function or the power spectrum

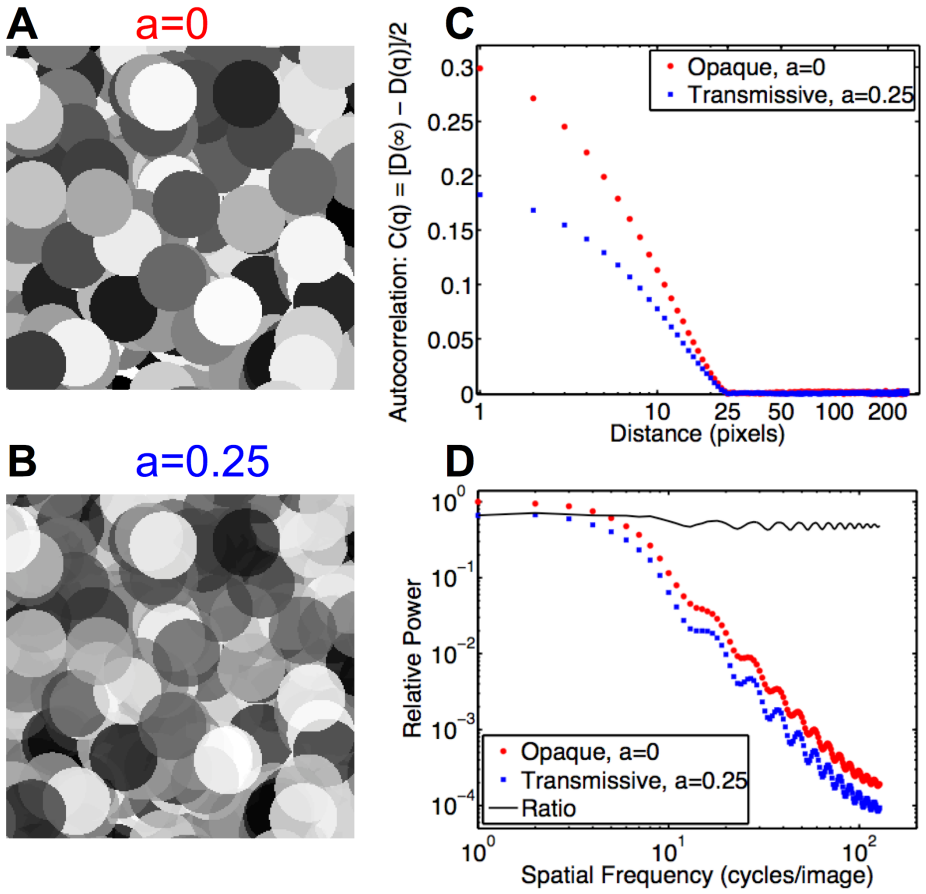

Figure 2.4: **For delta-function object size distributions, opaque and transmissive dead leaf images yield substantially different autocorrelation functions and power spectra.** (**A,B**) Sample images in which the leaves are all the same size ($s^\star = 25$ pixels), from the opaque (a=0, **A**) and transmissive (a=0.25, **B**) image ensembles. The autocorrelation functions of these image ensembles are shown in panel **C**. Their power spectra (**D**) differ non-trivially. To highlight this fact, we plot the ratio between the power spectra and note that it is not constant. The ripples in the power spectra are indicative of the fixed length scale in the images: they occur at multiples of the $256/25 \approx 10$ cycles/img fundamental frequency. The ripples in the ratio of the power spectra indicate that the modulation depth depends on the leaf opacity.

of images: it is manifest only by a multiplicative constant in the power-law functions. For other object size distributions, this is not necessarily the case. Occlusion is important for natural image formation, but it does not change the form of the power spectrum. Since images formed by opaque leaves that are all the same size

have oscillatory, non-power-law, power spectra (Fig. 2.4), and transmissive leaves can yield power law power spectra (Fig. 2.3), occlusion is likely not responsible for scale invariance of images. The universality of power law power spectra in both occlusive imaging environments, like natural photographic images, and transmissive ones, like mammography, is likely due to power-law object size distributions in both settings.

The work presented in this section was done in collaboration with David Pfau.

Chapter 3

Using the 2-point functions of natural images to understand the peripheral visual system

Natural images contain highly redundant information. Consider, for example, the images shown in Fig. 2.1. In those images, nearby pixels tend to be similar – bright pixels tend to be next to other bright pixels, and dark pixels tend to be next to other dark pixels – the brightness field is highly correlated. As I demonstrated in Ch. 2, the autocorrelation function that describes this phenomenon is somewhat universal for natural images, and can be modeled using collages of randomly placed objects. Indeed, natural movies are also highly redundant in the time domain [Dong and Atick, 1995a], a fact to which I will return later in this chapter.

3.1 Redundancy is information theoretically inefficient

To understand the impact of redundancy on coding efficiency, we must first summarize some basic results from information theory, originally due to [Shannon, 1948]. For a modern, pedagogical, discussion of information theory, interested readers should read the excellent book by Cover and Thomas [Cover and Thomas, 1991].

Consider a communication channel along which we send messages $\mathbf{x}$, with corresponding probabilities $p(\mathbf{x})$. The *entropy* of this message distribution, $H[\mathbf{x}]$, is given by

$$H[\mathbf{x}] = -\sum_i p(\mathbf{x}_i) \log_2 \left(p(\mathbf{x}_i) \right), \tag{3.1}$$

and is measured in units of *binary digits*, or "bits". The entropy is a measure of

how informative a given message from this distribution is expected to be. In the case that only one message ever occurs, $p(\mathbf{x}_i) = \delta_{i1}$, the distribution contains no information since one knows, before the message is received, exactly what it will contain. A uniform distribution over messages results in each message being maximally informative and, in this case, the entropy of the distribution is just the logarithm of the number of possible messages, similar to how, in statistical physics, the entropy of a given macro state is the logarithm of the number of different microstates that give rise to the same macroscopic observables.

In the case of information theory, one can show that the entropy described in Eq. 3.1 gives the minimum number of binary symbols needed to recode the messages [Cover and Thomas, 1991] or, similarly, the minimum capacity a noiseless digital channel requires in order to send the messages.

When two messages are sent simultaneously (say, the outputs $\mathbf{x}$ and $\mathbf{y}$ of neighboring retinal ganglion cells), the joint entropy is simply

$$H[\mathbf{x}, \mathbf{y}] = -\sum_{i,j} p(\mathbf{x}_i, \mathbf{y}_j) \log_2 \left(p(\mathbf{x}_i, \mathbf{y}_j) \right), \tag{3.2}$$

where $p(\mathbf{x}_i, \mathbf{y}_j)$ is the joint probability distribution over the messages $\mathbf{x}$ and $\mathbf{y}$. With a small amount of algebra, it is easy to show that [Cover and Thomas, 1991]

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] + H[\mathbf{y}] - \sum_{i,j} p(\mathbf{x}_i, \mathbf{y}_j) \log_2 \left(\frac{p(\mathbf{x}_i, \mathbf{y}_j)}{p(\mathbf{x}_i) p(\mathbf{y}_j)}) \right), \tag{3.3}$$

where the last term is called the *mutual information* between variables $\mathbf{x}$ and $\mathbf{y}$,

$$I\left[\mathbf{x}; \mathbf{y}\right] = \sum_{i,j} p(\mathbf{x}_i, \mathbf{y}_j) \log_2 \left(\frac{p(\mathbf{x}_i, \mathbf{y}_j)}{p(\mathbf{x}_i) p(\mathbf{y}_j)} \right). \tag{3.4}$$

$I\left[\mathbf{x}; \mathbf{y}\right]$ describes the extent to which knowledge of the value of one random variable, say $\mathbf{x}$ can provide information about another one, say $\mathbf{y}$. When the two variables are statistically independent, $p(\mathbf{x}_i, \mathbf{y}_j) = p(\mathbf{x}_i) p(\mathbf{y}_j)$, and the logarithm in Eq. 3.4 is zero.

From this discussion, it is clear that, in order for the retinal ganglion cell (RGC) outputs to collectively be maximally informative, they should be statistically independent, so that their mutual information is zero, and thus the entropy of their joint message distribution is maximized. This can be intuitively understood by noting that, if there is dependency between the RGC outputs, then the output of cell A, in addition to specifying the state of that cell, also specifies (at least partially) the state of cell B. But, since cell B is sending its own output as well, that information is already present, and thus one could send the message more cheaply by removing

the shared part of the message. Of course, if the communication channel is noisy, the redundancy can be desirable, as it provides the possibility for error correction in the received message [Mackay, 2003]. The notion that sensory systems should be information theoretically efficient is an old one, dating back to at least the 1960's [Barlow, 1961], and has had a significant impact on the field.

3.2 Whitening in the spatial domain: a model for RGC filtering?

Retinal ganglion cells tile the retina, with each RGC being loosely analogous to a pixel on a CCD chip, although the RGCs are in a hexagonal closest-packing configuration [Wassle *et al.*, 1981], and not arrayed on a square lattice like CCD pixels. Since natural images contain significant spatial correlations (Ch. 2), this means that nearby RGCs get similar input from the photoreceptors. From the discussion in the beginning of this chapter, recall that efficient coding arguments suggest independence of RGC activity.

To make progress on connecting efficient coding and RGC physiology, I will relax the requirements slightly, and focus on correlations rather than statistical dependency. There are several reasons do to this: experimentally, it requires much less data to measure pair-wise correlations than the 3-pt (and higher) functions; in the case of 2-point statistics, the convolution theorem provides a very nice connection between linear filter properties and image power spectra; in the case of jointly Gaussian variables (although RGC activity is not jointly Gaussian [Schneidman *et al.*, 2006]), removing correlations is equivalent to introducing full independence.

If one assumes that RGCs are linear filters, such that their outputs (for which I will use firing rates $r(\vec{x})$ – where $\vec{x}$ is the RGCs location – as a proxy) are given by the convolution between a filter $Q(\vec{x}-\vec{x}')$, and the input image with pixel values $I(\vec{x'})$, $r(\vec{x}) = \int d\vec{x'} Q(\vec{x}-\vec{x}')I(\vec{x}')$, and that the RGCs all have the same filter, the problem can be reduced to asking: which filter $Q(\vec{x}-\vec{x}')$, when convolved with natural images will yield spatially uncorrelated outputs [Atick and Redlich, 1992]?

Atick and Redlich provide more rigorous derivations, but one can quickly solve this problem by recalling that an uncorrelated signal has a delta function autocorrelation function, the Fourier transform of which is a constant. Recalling the convolution theorem, which states that the Fourier transform of a convolution is the product of the Fourier transforms, and that the amplitude spectra (modulus of Fourier transform) of natural images are $1/k$ for spatial frequency k, it becomes clear that the decorrelating filter should have an amplitude spectrum $\tilde{Q}(k) \propto k$ [Atick and Redlich, 1992].

In the presence of noise, this filter is not so useful, because natural images have very little power at high spatial frequencies (recall their k^{-2} power spectra), but any

additive noise, like shot noise, will tend to be spatially uncorrelated, and thus persists up to arbitrarily high spatial frequency: white noise has a frequency-independent power spectrum, hence the reason it is called "white". Thus, above some threshold spatial frequency, there is more noise power than signal power, and continuing to convolve the signal with a k filter past this frequency amplifies the noise to potentially disastrous levels. One might thus predict that the optimal RGC filter would be linear in spatial frequency for small k, followed by a roll-off (low pass filtering) above some threshold spatial frequency. Indeed, physiologically measured retinal receptive fields show precisely this trend [Atick and Redlich, 1992].

Thus, as far as linear filtering is concerned, decorrelation (whitening), with noise-reducing low-pass filtering, seems to provide a decent account of RGC physiology. Direct measurements show that the correlations between RGC spike trains are small, with the distribution of Pearson's linear correlation coefficients being sharply peaked near zero, and having an RMS value of much less than 0.1, although even these small correlations can have significant impacts on the collective network activity [Schneidman *et al.*, 2006].

Of course, one must not believe too strongly in the linear filtering arguments, since neuronal outputs are significantly non-linear functions of their inputs. Canonically, one considers a neurons as a linear-nonlinear system, in which a linear filter acts on the inputs to the cell – like, say, the stimulus convolved with a receptive field – , and the cell output is a pointwise nonlinear function of that convolution [Pillow *et al.*, 2008; Carandini *et al.*, 1997]. Interestingly, very recent work has shown that the nonlinear transfer properties of RGCs, rather than their linear filters, are primarily responsible for decorrelating their activities [Pitkow and Meister, 2012].

Regardless of the cause of the decorrelation, later in this thesis, I will use whitened natural images as inputs to a model of visual cortex, as is standard practice in the field [Simoncelli and Olshausen, 2001].

3.3 Lateral geniculate nucleus outputs are white in the time domain

While RGCs appear to perform spatial decorrelation of natural images, the temporal correlations in natural movies provide another, potentially significant, source of redundancy and / or inefficiency. Much like in the spatial domain (Ch. 2), natural movies have f^{-2} power spectra in the time domain, where f is the temporal frequency [Dong and Atick, 1995a]. Thus, linear filtering arguments similar to those used in the preceding section suggest that, somewhere in the peripheral visual system, there exist neurons whose response functions have f^2 power spectra.

To address this question, Yang Dan and colleagues recorded spike trains from the

lateral geniculate nucleus (LGN – the structure that receives visual input from the eyes via the optic nerve, and relays that information to the primary visual cortex, V1) of cats viewing various stimuli.

In response to natural movies (Casablanca was used as a stimulus), the LGN neurons' spike trains had flat power spectra, thus they were whitening the f^{-2} input [Dan *et al.*, 1996]. As a control, Dan *et al.* repeated their experiment with white noise stimuli, and found that the power spectra of the LGN responses were $\sim f^2$. In other words, LGN activity is consistent with applying a filter with an f amplitude spectrum (and thus f^2 power spectrum) to the input stimulus. When that stimulus is a natural movie with $1/f$ amplitude spectrum, the output is white, and when the input is white (with constant power spectrum), the output power spectrum varies as f^2.

Of course, it is not clear from this experiment alone whether the f filter is applied at the LGN, at the retina, or somehow through the concatenation of these two transformations. Other work shows that the LGN filters (spike triggered average temporal response) are similar to what one expects, if the LGN is performing the temporal whitening [Dong and Atick, 1995b].

Chapter 4

Characterizing the higher-order statistics of natural images

As I have discussed in the Chapters 2 and 3, natural images have nearly universal autocorrelation functions, and the peripheral stages of visual processing largely remove those correlations, resulting in "whitened" natural images (so-called because they have equal power at all frequencies, much like white light), an example of which is shown in Fig. 4.1. Compared to raw images, like those in Fig. 2.1, edges are much more apparent in the whitened image, as a result of the increased power at high spatial frequencies. That the whitened image contains significant structure (indeed, the semantic content of the image is still obvious, even after whitening) underscores the existence of significant higher-order correlations (beyond the 2-point function). In this chapter, I will summarize attempts to understand this higher-order structure, which will lead to interesting insights about the possible optimization principles underlying the physiology of visual cortex.

4.1 Independent Component Analysis

To build up a model for natural image statistics, imagine taking whitened images, and searching for a basis in which each dimension is statistically independent. Such an analysis would provide a model for the probability distribution function natural images, and the bases (features) so-found would inform us about the composition of natural images. This analysis is called Independent Component Analysis (ICA) [Bell and Sejnowski, 1997], and has had a profound effect on the computer vision community.

To perform ICA, one proposes that the image, which is reshaped into a vector $\vec{I}$, the elements of which represent the pixel values, is formed from a linear mixture of

Figure 4.1: **A whitened natural image.** Despite the absence of 2-point correlations, there is still much obvious structure in the image, suggesting a rich higher-order correlation function (beyond 2-point).

sources $\vec{A}$, via the mixing matrix Q: $\vec{I} = Q\vec{A}$. One then proposes a prior distribution over the source weights $\vec{A}$, and modifies the mixing matrix Q until the sources follow that prior distribution, which is typically a factorial distribution, $p(\vec{A}) = \prod_i \ p(A_i)$, such that the sources are independent. With reference to the discussion on redundancy and information theory in Ch. 3, one can see that, since the sources A are independent, one could form an efficient code for natural images by finding the source weights for each image, and transmitting those instead of the raw pixel values (which are somewhat redundant, even after the image has been whitened). This observation seeks to emphasize the fact that the most efficient way to encode a given class of signals depends strongly on the statistical properties of those signals and thus, by adapting to the stimulus statistics, sensory systems can improve their efficiency.

In the simplest form of ICA, the matrix Q is square and invertible, so the source weights can be easily determined as $\vec{A} = Q^{-1}\vec{I}$. It is then relatively straightforward to write down the image probability in terms of the prior over source weights, and to perform gradient ascent of the entries in the matrix Q on the log-likelihood function such that, on average, this probability is maximized. For the (very common) choice of a Laplace prior, $p(\vec{A}) = \prod_i \ e^{-|A_i|}$, this gradient ascent calculation yields the learning

rule

$$\Delta Q \propto \left\langle (Q^T)^{-1}\text{sign}(\vec{A})\vec{A}^T - (Q^T)^{-1} \right\rangle . \tag{4.1}$$

The specific details of Eq. 4.1 are not particularly important for this thesis, although I will later spend a lot of time and attention on other learning rules. Rather, this equation serves to emphasize the algorithm by which one performs ICA. First, one randomly initializes the matrix Q, so as to be agnostic to the structure of the independent components that are found. Then, one presents a set of images, the source weights for which are computed as $\vec{A} = Q^{-1}\vec{I}$. The matrix Q is then updated using the rule given in Eq. 4.1, with the constant of proportionality (called the "learning rate") being chosen such that the algorithm converges as quickly as possible. This process is then iterated until the matrix Q converges. Since the learning rule was derived as gradient ascent on the log-likelihood function, upon convergence, the matrix Q will maximize the likelihood of the image patches, under the prior distribution we have chosen. Of course, gradient ascent is prone to finding local maxima, and so there is no guarantee that this algorithm will find the global maximum for the matrix Q.

If the learning rate is chosen to be too large, the entries of Q will keep "bouncing around", taking large steps about the optimum, and will never converge. Conversely, if the learning rate is too low, the algorithm will be very slow to converge. To get around this, a common strategy is to start out with a large learning rate, and to gradually reduce it as the algorithm nears convergence.

Figure 4.2 shows an example of ICA basis functions trained on 16×16 pixel patches drawn randomly from whitened natural images, and trained to optimize a Laplace prior. Each of these basis functions corresponds to a column of the matrix Q, reshaped into a square image patch. One of the basis functions corresponds to the DC brightness level of the image patch, while the remainder are localized and oriented edge filters. A cursory comparison with the macaque V1 receptive fields in Fig. 1.5 shows some similarities, although the two feature sets are clearly not in quantitative agreement.

4.2 Sparse Coding

While ICA, as presented in the previous section, assumes that the image is a noiseless linear mixture of sources, and that number of sources is equal to the number of pixels, both of these assumptions can be relaxed, yielding the sparse coding analysis of [Olshausen and Field, 1996]. I will devote considerable attention to this, and related, previous works, because they form important background material for the novel work presented in Chapters 5 and 6.

In a sparse coding model, one recodes the image $\vec{I}$ in terms of the weights of a set

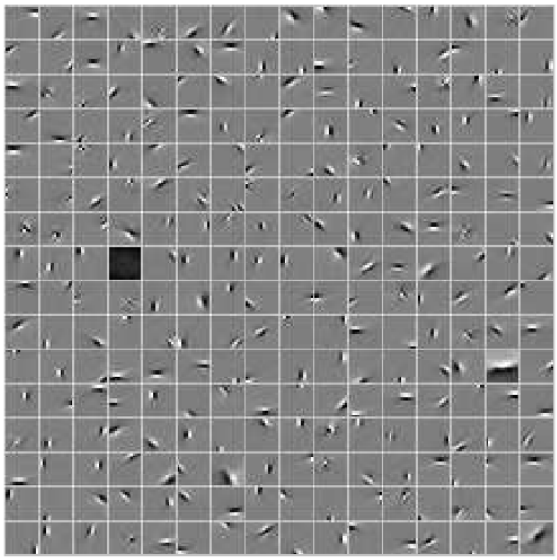

Figure 4.2: **ICA basis functions for whitened natural images.** Each square in the grid shows one ICA basis function for 16×16 pixel whitened natural image patches.

of basis functions (much like ICA) and postulates that the weights A_i of those basis functions in response to each image, and the basis functions themselves Q_{ij}, minimize the following cost function [Olshausen and Field, 1996]:

$$E = \frac{1}{2\sigma^2} \sum_j |I_j - \sum_i A_i Q_{ij}|^2 + \sum_i C(A_i), \tag{4.2}$$

where the first term is the RMS deviation between the image model (formed by multiplying each basis function by its weight and summing over all bases), the function $C(\cdot)$ is the sparseness penalty function, and σ is a constant that determines the relative importance of the two terms. Commonly the sparseness penalty function is the absolute value of the weights, $C(\cdot) = |\cdot|$, in which case the algorithm seeks to simultaneously minimize the error in the image model, and the so-called L_1 norm of the weights (the L_1 norm is simply the sum over the absolute values), although other penalty functions are possible.

Unlike ICA, the matrix Q need not be square, so the sparse code can be either *undercomplete* (having fewer bases than pixels), critically sampled, or *overcomplete*, in which case there are more bases than pixels.

Algorithmically, sparse coding works much like ICA: one starts with randomized bases, then determines the weights of those bases in response to whitened natural images. The bases are then updated to decrease the value of the cost function (gradient

descent).

Whereas in ICA, the source weights could be obtained by matrix inversion, the same is not true in sparse coding (SC) analyses. In the original paper on sparse coding, Olshausen and Field obtained the weights by gradient descent on their cost function [Olshausen and Field, 1996]. The gradient of the cost function with respect to the weights is

$$\frac{\partial E}{\partial A_i} = -2\left(\frac{1}{2\sigma^2}\right)\sum_j \left(I_j - \sum_k A_i k Q_{kj}\right) Q_{ij} + C'(A_i), \tag{4.3}$$

and thus gradient descent yields an update rule of

$$A_i(t+1) = A_i(t) + \eta\left[\frac{1}{2\sigma^2}\sum_j I_j Q_{ij} - \frac{1}{2\sigma^2}\sum_{j,k} A_k(t) Q_{kj} Q_{ij} - C'(A_i(t))\right], \tag{4.4}$$

where η is a small positive constant that determines the rate at which the weights change. To perform the inference – determine the weights corresponding to a given image – the weights A_i are initialized at zero, and the updates defined in Eq.4.4 are performed iteratively until convergence is achieved. More recently, Rozell and colleagues showed that this inference step could be implemented by a network of interconnected model neurons [Rozell *et al.*, 2008].

As in ICA, once the weights are known for a set of images, the bases are updated, using gradient descent on the cost function

$$\begin{aligned} \frac{\partial E}{\partial Q_{ij}} &= -2\left(\frac{1}{2\sigma^2}\right) A_i \left(I_j - \sum_k A_k Q_{kj}\right) \\ \implies Q_{ij} &\rightarrow Q_{ij} + \gamma A_i \left(I_j - \sum_k A_k Q_{kj}\right), \end{aligned} \tag{4.5}$$

where γ is a small positive constant. This interleaving of inference and learning is then repeated until the bases converge. We note that, in both sparse coding, and ICA, the basis functions are forced to have unit vector length, such that $\sum_j Q_{ij}^2 = 1\ \forall i$. This is implemented by normalizing the bases after each update. Without this requirement, the cost function (Eq. 4.2) could be trivially minimized by having large magnitudes on the bases, resulting in very small weights A_i.

The connection between ICA and SC can be understood by noting that the cost function in Eq. 4.2 can be interpreted as a negative-log-likelihood [Olshausen and

Field, 1997], for which the corresponding likelihood function is

$$p(\vec{I}) \propto \exp\left[-\frac{\sum_j |I_j - \sum_i A_i Q_{ij}|^2}{2\sigma^2}\right] \prod_i \exp(-C(A_i)). \tag{4.6}$$

In other words, sparse coding assumes that the true image is Gaussian distributed with respect to the image model $\sum_i A_i Q_{ij}$, and that the statistical prior on the weights is factorial.

A common choice for the sparse cost function is the so-called L_1 norm of the weights, in which case $C(A_i) = |A_i|$. In the original sparse coding paper, a slightly undercomplete (with respect to the number of pixels) representation was used, and, after training on natural images, both the L_1 sparseness penalty, and the use of a Cauchy prior resulted in basis functions that look like localized, oriented Gabor wavelets (somewhat like the ICA bases in Fig. 4.2). It was argued at the time [Olshausen and Field, 1997] that the agreement between these basis functions, and the receptive fields (RFs) of visual cortical neurons, might indicate that V1 forms a sparse code for natural images, although subsequent work [Ringach, 2002] showed that there were some discrepancies between the RF shapes and the sparse coded basis functions. For example, with respect to the macaque V1 RFs in Fig. 1.5, the sparse coding model [Olshausen and Field, 1996] failed to display the small unoriented "blobs".

More recently, Rehn and Sommer showed that a highly overcomplete sparse coding model with 3 times more bases than image pixels, which uses an L_0 sparseness penalty, could yield quantitative agreement between the basis function shapes, and the physiologically recorded V1 receptive fields [Rehn and Sommer, 2007]. The L_0 norm of a vector is simply the number of non-zero entries in that vector, and thus the model of Rehn and Sommer seeks to minimize the number of active units.

The above discussion begs the question: is it the correct choice of sparseness penalty (L_0 vs L_1 for example), or the overcompleteness level, that dictates whether a sparse coding algorithm will yield the full diversity of RF shapes seen in visual cortex? Given that massively overcomplete SC models that minimize the L_1 norm can also yield the full set of RF shapes [Olshausen *et al.*, 2009], I suspect that it is the overcompleteness, and not the specific choice of cost function, that is more important.

Intriguingly, theoretical work in statistics indicates that, if the image was truly L_0 sparse (meaning that, of a large dictionary of basis functions, the image could be reconstructed exactly by taking a weighted sum of no more than k of those bases, for some integer k), then finding the L_1-minimizing set of weights would give the same answer as seeking out the L_0-minimizing set of weights [Donoho, 2004]. Since the rule for updating the basis functions (Eq. 4.5) depends only on the weights, pixel values, and basis functions, and *not* explicitly on the cost function itself, this means that, in the case that the images are really L_0 sparse, L_1 and L_0 sparseness penalty

functions would be equivalent with regards to sparse coding models. However, the reader should note that this condition (k sparseness) is not necessarily satisfied by the image coding problem.

4.3 Might primary visual cortex use a sparse code for natural images?

As I have discussed in the above section, a sufficiently overcomplete sparse coding model learns, when trained on natural images, the same set of features that are displayed in the receptive fields of visual cortical neurons. This has prompted some to suggest that V1 might form an overcomplete, sparse representation [Olshausen and Field, 1997].

Aside from the receptive field shapes themselves, the specific sparse coding models(s) discussed above implies several testable predictions:

- V1 must be overcomplete with respect to the RGCs;
- cortical neurons should have sparse activity levels;
- cortical neurons should be statistically independent;
- one should be able to linearly decode the cortical activity, to recover the stimuli.

I note for sake of completeness that these last two predictions are not strictly necessary for *all* sparse coding models. Indeed, one can create sparse coding models in which the prior on neuronal activations is not factorial (thus not being statistically independent), or in which the "representation" is non-linear [Karklin and Simoncelli, 2011].

In the remainder of section, I will address these predictions one-by-one, and come to the conclusion that V1 appears to largely satisfy the above conditions.

4.3.1 Is V1 overcomplete with respect to the retinal ganglion cells?

The human primary visual cortex contains around 10^8 neurons [Leuba and Kraftsik, 1994; Klekamp *et al.*, 1991]. In contrast, there are around 10^6 retinal ganglion cells whose axonal projections form the human optic nerve [Curcio and Allen, 1990]. Thus, human V1 is ~ 100 times overcomplete with respect to its (retinal) input. Not all of the V1 neurons are expected to be involved in forming the sparse code described above. For example, V1 contains neurons that are tuned to retinal disparities, and

are thus involved in estimating depth, interneurons that are not suspected to be involved in this coding, and complex cells that respond to specific image features in a translation-invariant manner. Nevertheless, V1 is sufficiently overcomplete that a subset of its neurons (the so-called "simple cells") could be involved in the formation of such a code.

As a counterargument, recall that visual cortical neurons respond different to repeated presentations of the same stimulus. It has been argued that, in order to achieve a reliable rate code, one must average over $50 - 100$ different neurons [Shadlen and Newsome, 1994]. If Shadlen and Newsome are correct, then the observed overcompleteness might not be related to sparse coding, but would rather exist to introduce enough redundancy in visual cortex to allow it to operate reliably in the presence of noise, although other visual cortical studies [Mainen and Sejnowski, 1995] indicate that V1 neurons might be more reliable (less noisy) than they are often given credit for.

4.3.2 Do (visual) cortical neurons have sparse activity levels?

Measurements of the firing rates of V1 neurons in response to videos of natural scenes show that those rates are low, and that the firing rate distributions are sharply peaked near zero [Baddeley *et al.*, 1997], and follow a roughly exponential distribution, similar to what one finds in sparse coding models that minimize the L_1 sparseness cost function. Similarly, cell-attached recordings in auditory cortex show highly sparse levels of activity [Hromádka *et al.*, 2008], with the distribution of neuronal firing rates following a roughly lognormal function. The comparison between auditory and visual cortices might at first seem a bit odd, but it is generally assumed (or at least hoped) that, on some level, "cortex is cortex", and one can generalize between different cortical areas.

Conversely, other experimenters [Tolhurst *et al.*, 2009] have observed non-sparse (dense) neuronal activity in visual cortex, although the boundary between "sparse" and "dense" activity is open to interpretation and thus it is unclear how sparse the activity must be in order to confirm the sparse coding hypothesis [Zylberberg *et al.*, 2011].

Interestingly, however, the relative level of sparseness of cortical responses to natural images increases when a larger fraction of the visual field is covered by the stimulus [Vinje and Gallant, 2002; Vinje and Gallant, 2000; Haider *et al.*, 2010], as a result of inhibitory interneuronal connections [Haider *et al.*, 2010].

With respect to direct measures of sparseness, then, I conclude that there is some ambiguity, but the balance of evidence does not rule out a sparse code.

Intriguingly, sparseness levels in ferret V1 appear to *decrease* during maturation: as the animal gains more visual experience, its visual cortical activity becomes less

sparse [Berkes *et al.*, 2009]. This is counter to what one might expect from the sparse coding hypothesis, and I will address this observation in detail in Ch. 6.

4.3.3 Do visual cortical neurons have statistically independent activities?

As I have alluded to previously, it is very difficult to experimentally assess statistical independence, because a very large amount of data is required in order to fully sample the joint probability distribution function, which is needed in order to make that claim. More readily accessible are the pair-wise correlations, which must be zero if the neurons are independent (although the absence of correlations is not sufficient to prove independence), and which can be measured with a relatively modest amount of data. Recent experimental work [Ecker *et al.*, 2010; Renart *et al.*, 2010] has shown that neurons in visual cortex tend to have very small correlations between their firing rates: the mean ($\pm$ SEM) Pearson's linear correlation coefficients between visual cortical neurons is 0.01 ± 0.002, and the correlations are not significantly larger for neurons with similar receptive fields than for neurons with very different receptive fields [Ecker *et al.*, 2010].

The reader should note that these results are highly controversial, with previous studies reporting much larger average correlation coefficients, some of which were larger than 0.3. [Ecker *et al.*, 2010; Gawne *et al.*, 1996; Reich *et al.*, 2001; Kohn and Smith, 2005; Gutnisky and Dragoi, 2008; Smith and Kohn, 2008]. Ecker et al. claim that this large discrepancy is a result of potentially significant systematic errors in the other groups' results [Ecker *et al.*, 2010].

As an example, when one performs extracellular recordings (which is the norm when recording from several neurons at once), multiple, very fine, and very near to each other, wires are inserted into the cortex, and voltage traces are recorded. These are then high-pass filtered and thresholded to obtain spikes. The remaining problem is to identify which spikes came from which neurons, and this is no easy task. Since different neurons will be different distances from the electrodes, and have different orientations with respect to the electrodes, their spikes can yield different voltage traces across the set of electrodes. One then attempts to cluster the spikes into groups, with each cluster corresponding to a (putative) single neuron. Of course, if some of the spikes are mis-assigned, then the apparent correlations could be poorly estimated. This is less of an issue with retinal recordings, like those of [Schneidman *et al.*, 2006], where a piece of disembodied retina is laid directly on top of the recording electrodes, so the physical locations of the retinal ganglion cells (which tile the surface of the retina) correspond more closely to the electrode locations. Thus, the retinal correlation studies are less controversial.

Similar to the sparseness measurements discussed in the previous section, the

correlations present in cortical responses to natural images decrease when a larger fraction of the visual field is covered by the stimulus [Vinje and Gallant, 2002; Vinje and Gallant, 2000; Haider *et al.*, 2010], as a result of inhibitory interneuronal connections [Haider *et al.*, 2010].

As in the case of the direct sparseness measurements, I conclude that V1 has mechanisms that decorrelate its neuronal activities (namely, lateral inhibition [Haider *et al.*, 2010]), and that the resultant correlations may be small, but this is still an active area of research.

Recently, it has been observed that the statistical dependence between neurons increases during development. In particular, the Kullback-Leibler (KL) divergence between the probability distribution function (PDF) of neural activities, and the product of the marginal distributions over each neuron's state (which assumes independence) increases during development [Berkes *et al.*, 2009; Berkes *et al.*, 2011]. The KL divergence measures how different two probability distributions are from each other. This observation (and others) have been interpreted to mean that cortex is not explicitly forming a sparse representation, but rather that V1 learns a Bayesian prior over image statistics [Berkes *et al.*, 2011].

Thus, there is much controversy over whether or not V1 neuronal activities are statistically independent. As we noted above, one could create a sparse coding model with a non-factorial prior over neuronal activations, and thus the independence property is not strictly necessary for sparse coding models.

4.3.4 Can cortical activity be linearly decoded to recover the stimulus?

This prediction is by far the least experimentally accessible of the ones listed above. Given the (enormous) number of neurons in V1, answering this question is quite a challenge. The challenge is reduced somewhat by the topographic organization of V1, wherein the physical layout of V1 forms a (distorted) map of the retina [Dayan and Abbott, 2001; Adams and Horton, 2002; Engel *et al.*, 1997]. Put another way, nearby neurons in V1 have receptive fields located in similar portions of visual space, and the organization of these receptive fields across the surface of V1 forms a map of the retina.

Thus, if one could record from a sufficiently large number of very nearby neurons, one could ask if this sub-region of V1 can be linearly decoded to recover the portion of visual input from the corresponding region of visual space. Unfortunately, even this simplified study is still extremely difficult, and thus it has not been performed.

Assuming that some aspects of neurophysiology, and neural computation, generalize across neural systems, however, some insight might be gained from similar studies performed on other preparations. For example, Stanley and colleagues recorded si-

multaneous activity from 177 neurons in the cat LGN, and trained a linear decoder to estimate the stimulus from these responses. On test data (not used to train the decoder) the linear decoding could account for $20\% - 50\%$ of the variance in the input stimulus, and the decoded estimate contained many qualitative features of the stimulus [Stanley *et al.*, 1999] . While these results are not quantitatively that compelling, 177 is a very small fraction of the number of neurons in LGN, and so a better decoding is almost certainly possible.

Similarly, Bialek and colleagues recorded from the H1 motion-sensitive neurons in a blowfly in response to a randomly moving stimulus, and then found the optimal (causal) linear decoder to recover the velocity profile (velocity as a function of time) from the spiking activity [Bialek *et al.*, 1991]. They found that the linear decoder could do a reasonably good job of recovering the stimulus (on held-out test data not used to train the decoder), and that the addition of quadratic terms in their decoder – corresponding to the second-order terms in the Volterra expansion – did not significantly improve the decoding accuracy.

I conclude that some neural systems have been observed to display linear decodability, but that this has yet to be studied in the visual cortex. One can also generalize the notion of sparse coding beyond linear image models. For example, Karklin and Simoncelli recently published a (non-linear) model, in which non-linear functions of the image were learned so as to maximize the mutual information between the function outputs and the image input, while maintaining sparseness of the outputs [Karklin and Simoncelli, 2011]. Thus, it may not be necessary for neural systems to be linearly decodable, even if they are performing some variant of sparse coding.

4.4 Summary

To recap, even whitened images contain significant redundancy, which can be reduced by an appropriate transformation (recoding) of the image. One popular proposal is that this is done by sparse coding – somewhat analogous to regularized regression – in which weights and basis functions are chosen such that a linear reconstruction of the image is possible, using minimal values for the weights.

This is an appealing proposal for neural systems, because this notion of minimizing the weights would correspond to minimizing neuronal activity, which leads to energetic efficiency. At the same time, sparse coding is an appealing process because it helps to find parsimonious "explanations" for the stimulus in terms of environmental causes [Olshausen and Field, 1997; Rehn and Sommer, 2007]. I will discuss this process briefly below, but first note that this is valuable because, at some level in the hierarchy, the visual system contains a semantic description of the environment, in which discrete objects are identified, and the transformation from raw image to

such a high-level description is facilitated by parsimonious causal explanations of the stimulus.

I conclude from this discussion that sparse coding is an appealing proposal, and that there is some experimental support, of varying degrees of strength, to suggest that it may, in fact, provide a reasonable systems-level explanation for many physiological properties of the primary visual cortex.

4.4.1 Explaining Away

I will finish this chapter with a brief discussion of the notion of "explaining away", which helps to emphasize some of the power of sparse coding for image processing [Olshausen and Field, 1997; Rehn and Sommer, 2007].

Consider a large dictionary of basis functions, some subset of which are active in response to any given image. If the dictionary is overcomplete, then the bases cannot by mutually orthogonal, and thus some will have (potentially large) overlap with each other. In a purely feed-forward network (with no lateral interactions), the weights of the bases will depend only on the projection of the image onto said bases. Thus, similar (strongly overlapping) bases will have similar weights in response to a given image. Since the bases are (by assumption) not identical, one of them will have a larger projection onto the image, and will thus provide a better match for the image features. In seeking a parsimonious "explanation" for the image, then, it would be desirable to suppress the feature with less-than-maximal projection onto the image, in the case where a similar feature is more strongly represented. This suppression is called "explaining away", and can be implemented in a network model of sparse coding by lateral interactions between the units [Rozell *et al.*, 2008; Rehn and Sommer, 2007]. Similarly, the lateral inhibition in visual cortex could be implementing this idea [Haider *et al.*, 2010; Vinje and Gallant, 2002; Vinje and Gallant, 2000].

Chapter 5

(How) can biophysical neurons learn a sparse code for natural images?

In the previous chapter, I argued that primary visual cortex really does appear to form a sparse code for natural images. However, given that the visual cortical receptive fields are a product of development (and not hard-coded from birth) [Imbert and Buisseret, 1975; Cynader *et al.*, 1973], one should not be completely satisfied with the sparse coding proposal until it is understood *how* it might be implemented by biological systems, as well as *why* it is a good idea, and *what* its observable consequences might be.

With regards to these last two points, I refer the reader to the previous chapter, which summarizes the extensive literature on this topic. As I will demonstrate in this chapter, however, the existing work fails to address the question of *how* a sparse code might be learned by a group of biophysical neurons (or, indeed, whether this is even possible), and thus cannot be completely satisfactory in and of itself.

5.1 Sparseness and decorrelation allow biophysical neurons learn a linear generative code for natural images

Consider the learning rule via which the sparse coding algorithms of [Olshausen and Field, 1996; Rehn and Sommer, 2007] update their bases (Eq. 4.5), and which optimizes the linear generative model formed by the weights (A_i) of the bases:

$$\Delta Q_{ij} \propto A_i \left(I_j - \sum_k A_k Q_{kj} \right). \tag{5.1}$$

Since we generally use these bases to attempt to understand the receptive fields of V1 neurons, and those receptive fields (presumably) describe the regions in visual space from which the cell receives synaptic input, we would like to consider the values Q_{ij} as the strengths of synaptic connections between V1 neurons (indexed by i), and LGN neurons (which relay RGC outputs – roughly analogous to image pixels, indexed by i). Indeed, as I will show later in this chapter, even when the neurons are not linear time-invariant filters (as in Ch. 1.3.3), the spike-triggered average stimulus (STA) can still recover exactly the strengths of the inputs from image pixels to the neurons.

However, once one makes the connection between the bases and synaptic strengths, it is clear that the update (learning) rule described in Eq. 5.1 is not biologically plausible because it is synaptically non-local. Recall that synaptic strengths in cortex are modified depending on the pre- and post-synaptic activities [Feldman, 2009; Abbott and Nelson, 2000], which in this case correspond to the inputs I_j, and the outputs A_i.

In Eq. 5.1, however, updates to the "synaptic" connection Q_{ij} depend on all the activities $\{A_i\}$ of all of the neurons in the network, while it is not clear that each individual synapse would have access to that information.

One subtle point, which does not affect my argument (above), is that of synaptic scaling. Indeed, it appears that cortical neurons can simultaneously scale all of their synaptic strengths either up or down by the same amount, so as to homeostatically regulate the amount of input they receive [Abbott and Nelson, 2000]. This is roughly analogous to having the update to Q_{ij} depend on $[\{Q_{ij}\}\ \forall j]$, in addition to A_i and I_j, in which the update to any particular synapse on a given cell depends on what happens at other synapses on the same cell. This mechanism does not allow, however, for updates to a synapse to depend directly on the activities of other cells that do not connect at that synapse.

In the remainder of this section, I will discuss a resolution to this problem that allows for synaptically local plasticity rules to result in the formation of an optimal linear generative image model by the neuronal activities [Zylberberg *et al.*, 2011].

Consider the non-local update rule in Eq 5.1. Expanding the polynomial, and averaging over image presentations,

$$\langle \Delta Q_{ij} \rangle \quad \propto \quad \langle A_i I_j \rangle - \left\langle A_i^2 Q_{ij} \right\rangle - \sum_{k \neq i} \langle A_i A_k Q_{kj} \rangle \,. \tag{5.2}$$

If the learning rate is small, such that the feed-forward weights change only slowly over time, then one can approximate that they are constant over some (small) number of image presentations, and take them outside of the averaging brackets:

$$\langle \Delta Q_{ij} \rangle \quad \sim \quad \langle A_i I_j \rangle - \left\langle A_i^2 \right\rangle Q_{ij} - \sum_{k \neq i} \langle A_i A_k \rangle Q_{kj}. \tag{5.3}$$

Now, so long as the neuronal activities are uncorrelated, and all units have the same average activity level μ (when averaged over many image presentations; I will later discuss how these constraints might be enforced, and whether or not they are realistic), $\langle A_i A_k \rangle = \langle A_i \rangle \langle A_k \rangle = \mu^2 \; \forall i, k$, and thus the learning rule reduces to

$$\langle \Delta Q_{ij} \rangle \quad \sim \quad \langle A_i I_j \rangle - \left\langle A_i^2 \right\rangle Q_{ij} - \mu^2 \sum_{k \neq i} Q_{kj}. \tag{5.4}$$

This last term could be small compared to the first two for a few reasons (which are easily satisfied). First, if the neurons in the network have *sparse* activity, meaning they are highly selective to particular image features, then $\langle A_i^2 \rangle \gg \langle A_i \rangle^2 = \mu^2$, and the second-to-last term dwarfs the last one.

Furthermore, the last term, $\mu^2 \sum_{k \neq i} Q_{kj}$, involves a sum over many different basis functions (model neuronal receptive fields). Some of these will be positive for a given pixel, whereas others will be negative. These random signs mean that the sum tends towards zero.

Thus, in the limit of sparse and uncorrelated neuronal activity, gradient descent on the error function E yields approximately

$$\langle \Delta Q_{ij} \rangle \quad \sim \quad \langle A_i I_j \rangle - \left\langle A_i^2 \right\rangle Q_{ij}, \tag{5.5}$$

which is equivalent to the average update from Oja's implementation of Hebbian learning, which was originally proposed as a way for a single neuron to learn the first principal component of the stimulus [Oja, 1982]. This rule is synaptically local inasmuch as the change in the connection strength between pixel j and neuron i depends only on the pre-synaptic activity I_j, the post-synaptic activity A_i, and the current value of the connection strength Q_{ij}.

Thus, a neuronal network *could* learn to approximately solve the same error minimization problem as did previous, non-local sparse coding algorithms [Rehn and Sommer, 2007; Olshausen and Field, 1996], using only synaptically local plasticity rules.

Interestingly, this result suggests that, even if the model neurons' outputs (say, spikes) are generated from the input in a highly non-linear way, if the network uses the above learning rule (Eq 5.5), then a linear decoding of the network activity should

provide a good match to the input: $I_j \approx \sum_i A_i Q_{ij}$. This linear decodability has previously been observed in physiology experiments [Bialek *et al.*, 1991; Stanley *et al.*, 1999], as well as models designed to maximize the information rate about input stimulus conveyed by individual spiking neurons [Bialek *et al.*, 1993; DeWeese, 1996].

I conclude this section by reiterating that, if the neuronal activities can be made sparse and uncorrelated (as they appear to be in cortex [Haider *et al.*, 2010; Vinje and Gallant, 2002; Vinje and Gallant, 2000; Hromádka *et al.*, 2008; Baddeley *et al.*, 1997]), then a synaptically local plasticity rule can suffice to learn to form an optimal linear generative model of the input. The remainder of this chapter will discuss the construction of a model neuronal network that does precisely this, and the study of the properties of that network.

5.2 The Sparse And Independent Local network (SAILnet)

5.2.1 SAILnet architecture and dynamics

Previous work has shown that the inference step (determining the maximum-lihelihood source weights A_i) in a sparse coding model can be implemented by a network of neurons with both lateral (between these neurons), and feed-forward (from the image pixels, in a rough approximation of the thalamic input to V1) connections [Rozell *et al.*, 2008]. The nature of the sparseness penalty (L_0 or L_1 norm for example) depends on the thresholding function in the model neuron. When integrate-and-fire dynamics are used to generate the neuronal activities, these networks minimize the L_0 cost function [Shapero *et al.*, 2011]. Such an architecture is described in Fig. 5.1.

Independent of this motivation, I use leaky integrate-and-fire (LIF) neurons as a biophysically realistic, yet simple, way to generate model neuronal activities in response to input stimuli. My main interest is in the learning process, and the specifics of how the dynamics perform inference are not crucial to this work, although they are an interesting and active area of research.

Towards this end, I implement a network of spiking, leaky integrate-and-fire units [Dayan and Abbott, 2001; Izhikevich, 2007] as model neurons. As in many previous models [Rozell *et al.*, 2008; Perrinet *et al.*, 2004; Falconbridge *et al.*, 2006; Földiák, 1990; Hopfield, 1984], each unit has a time dependent internal variable $u_i(t)$ and an output $y_i(t)$ associated with it. The simulation of my network operates in discrete time. The neuronal output at time t, $y_i(t)$, is binary-valued: it is either 1 (spike) or 0 (no spike), whereas the internal variable $u_i(t)$ is a continuous-valued function of time that is analogous to the membrane potential of a neuron. When this internal variable exceeds a threshold θ_i, the unit fires a punctate spike of output activity that

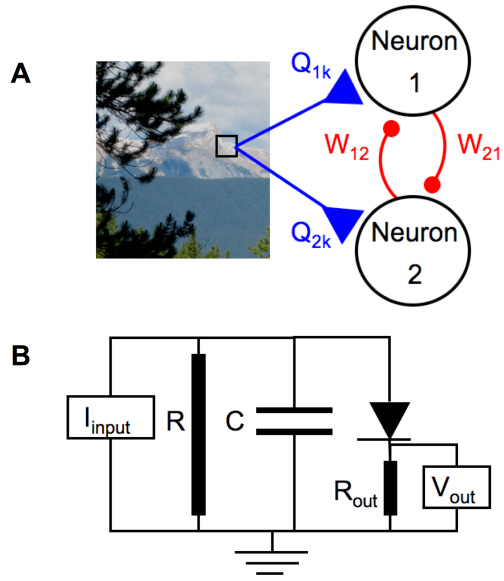

Figure 5.1: **SAILnet network architecture and neuron model** (A) The network architecture is based on that of Rozell, and inspired by recent physiology experiments. Inputs I_j to the network (from image pixels) contact the neuron at connections (synapses) with strengths Q_{ij}, whereas inhibitory recurrent connections between neurons in the network have strengths W_{ik}. The outputs of the neurons are given by $y_i(t)$; these spiking outputs are communicated through the recurrent connections, and also on to subsequent stages of sensory processing, such as cortical area V2, which are not included in this model. (B) Circuit diagram of the simplified leaky integrate-and-fire neuron model. The inputs from the stimulus with pixel values I_j, and the other neurons in the network, combine to form the input current $\mathcal{I}_{input}(t) = \sum_j Q_{ij} I_j - \sum_{k \neq i} W_{ik} y_k(t)$ to the cell. This current charges up the capacitor, while some current can leak to ground through a resistor in parallel with the capacitor. The resistors are shown as cylinders to highlight the fact that they model the collective action of ion channels in the cell membrane. The internal variable evolves in time via the differential equation for voltage across our capacitor, in response to input current $\mathcal{I}_{input}$: $du_i(t)/dt + u_i(t) = \mathcal{I}_{input}(t)$, which I simulate in discrete time. Once that voltage exceeds threshold θ_i, the diode, which models neuronal voltage-gated ion channels, opens, causing the cell to fire a punctate action potential, or spike, of activity. For sake of a complete circuit diagram, the output is denoted as the voltage, V_{out}, across some (small: $R_{out} \ll R$) resistance. After spiking, the unit's internal variable returns to the resting value of 0, from whence it can again be charged up.

lasts for one time step. This thresholding feature plays the role of neuronal voltage-gated ion channels (represented, as in Hopfield's circuit model [Hopfield, 1984], by a

diode) whose opening allows cortical neurons to fire. Other units in the network, and the inputs I_j, which are pixel intensities in an image, modify the internal variable $u_i(t)$ by injecting current into the model neuron. The structure of my network, and circuit diagram of the neuron model, are illustrated in Fig. 5.1, as are the model dynamics.

Similar to V1 [Haider *et al.*, 2010; Vinje and Gallant, 2002; Vinje and Gallant, 2000], inhibitory lateral connections (with strengths W_{ik}) will act to decorrelate the model neurons. I will further enforce sparseness by allowing each neuron's firing threshold (θ_i) to vary slowly over time, where the variation in firing threshold will seek to maintain a target long-term-averaged firing rate. This variable threshold mimics physiological homeostatic mechanisms like the variable intrinsic excitability of cortical neurons [Turrigiano *et al.*, 1998], or synaptic scaling [Abbott and Nelson, 2000; Marder and Goaillard, 2006] (discussed briefly in the previous section).

For the results presented in this chapter, I will study a network of $256 \times 6 = 1536$ neurons that recode 16×16 pixel whitened image patches. The network is six-times overcomplete with respect to the number of input pixels. This mimics the anatomical fact that V1 contains many more neurons than does LGN, from which it receives its inputs. Owing to the computational complexity of the problem — there are $\mathcal{O}(N^2)$ parameters to be learned in a SAILnet model containing N neurons — it is prohibitively time consuming to consider networks that are much more than $6\times$ overcomplete.

This six-times overcompleteness is in a sense analogous to the three-times overcompleteness of the SSC network described by Rehn and Sommer, since the outputs of their computational units could be either positive or negative [Rehn and Sommer, 2007], while my model neurons can output only one type of spike. Thus, each of their units can be thought of as representing a pair of my model neurons, with opposite-signed receptive fields.

5.2.2 SAILnet plasticity rules and objective function

I assess the computational output of each neuron in response to a stimulus image I by counting the number of spikes emitted by that neuron, $A_i = \sum_t y_i(t)$, following stimulus onset for a brief period of time lasting five times the time constant τ_{RC} of the RC circuit. My simulation updates the membrane potential every 0.1 τ_{RC} , thus there are 50 steps in the numerical integration following each stimulus presentation. Consequently, at least in principle, 50 is the maximum number of spikes I could observe from one neuron in response to any image. Instead of counting spikes, one could use first-spike latencies to measure the computational output [VanRullen and Thorpe, 2002; Delorme *et al.*, 2000]; these two measures are highly correlated in my simulations, with shorter latencies corresponding to greater spike counts (data not

shown). The network learns via rules similar to those of Földiák [Földiák, 1990; Falconbridge *et al.*, 2006]. These rules drive each unit to be active for only a small but non-zero fraction of the time (lifetime sparseness) and to maintain uncorrelated activity with respect to all other units in the network:

$$\begin{aligned} \Delta W_{ik} &= \alpha(A_i A_m - \mu^2) \\ \Delta Q_{ij} &= \beta A_i (I_j - A_i Q_{ij}) \\ \Delta \theta_i &= \gamma(A_i - \mu), \end{aligned} \tag{5.6}$$

where μ is the target average value for the number of spikes per image, which defines each neuron's lifetime sparseness, and α, β, and γ are learning rates — small positive constants that determine how quickly the network modifies itself. Updating the feed-forward weights Q_{ij} in my model is achieved with Oja's implementation of Hebb's rule [Oja, 1982]; this rule is what drives the network to represent the input. Note that because the firing rates are low ($\mu = 0.05$ spikes per neuron per image, for the results shown in this chapter), and spikes can only be emitted in integer units, my model implicitly allows only small numbers of neurons to be active at any given time (so called "hard" sparseness, or L_0 sparseness), similar to what is achieved by other means in some recent non-spiking sparse coding models [Rehn and Sommer, 2007; Rozell *et al.*, 2008].

Unlike previous work [Olshausen and Field, 1996; Rehn and Sommer, 2007], which performed unconstrained optimization on a cost function penalizing both reconstruction error and network activity, my model's learning rules (above) can be viewed as a gradient descent approach to a *constrained* optimization problem in which the network seeks to minimize the error between the input pixel values $\{I_j\}$, and a linear generative model formed by all of the neurons $\overline{I_j} = \sum_i A_i Q_{ij}$, while maintaining fixed average firing rates and no firing rate correlations.

Given the neuronal activities A_i in response to an image, and their feed-forward weights Q_{ij}, one can form a linear generative model $\overline{I}$ of the input stimulus $\overline{I_j} = \sum_i A_i Q_{ij}$. The mean squared error between that model $\overline{I}$ and the true input I is $E = \sum_j \left(I_j - \sum_i A_i Q_{ij}\right)^2$, and the creation of a high fidelity representation suggests that this error function E, or one like it, be minimized by the learning process.

Suppose that the neuronal network is not free to choose any solution to this problem; instead it must satisfy constraints that require the neurons to have a fixed average firing rate of μ spikes per image, and minimal correlation between neurons. Indeed, neurons tend to have low mean firing rates when averaged across many different images, and those firing rates span a finite range of values [Hromádka *et al.*, 2008; Baddeley *et al.*, 1997; Abeles *et al.*, 1990], motivating this first constraint. The second constraint is justified by observations that neural systems tend to exhibit little

or no correlation between pairs of units [Ecker *et al.*, 2010; Renart *et al.*, 2010], and that the correlation between the activity of V1 neurons decreases significantly as one increases the fraction of the visual field that is stimulated [Vinje and Gallant, 2000].

I use the method of Lagrange multipliers to solve this problem, allowing the learning rules to adapt the network so as to minimize reconstruction error while approximately satisfying these constraints. To do this, I perform gradient descent on a Lagrange function $\mathcal{L}$ that contains both the error function and the constraints:

$$\begin{aligned} \mathcal{L} &= \sum_j \left(I_j - \sum_i A_i Q_{ij} \right)^2 \\ &+ \sum_i \lambda_i (A_i - \mu) + \sum_{i \neq k} \tau_{ik} (A_i A_k - \mu^2), \end{aligned} \tag{5.7}$$

where the sets of values $\{\lambda_i\}$ and $\{\tau_{ik}\}$ are the (unknown) Lagrange multipliers. To perform constrained optimization, gradient descent is performed with respect to all of the free parameters in $\mathcal{L}$: namely, the set of feed-forward connection strengths $\{Q_{ik}\}$, and the Lagrange multipliers $\{\lambda_i\}$ and $\{\tau_{ik}\}$:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \lambda_i} &= A_i - \mu \\ \frac{\partial \mathcal{L}}{\partial \tau_{ik}} &= A_i A_k - \mu^2 \\ \frac{\partial \mathcal{L}}{\partial Q_{ij}} &= -2 A_i (I_j - \sum_k A_k Q_{kj}). \end{aligned} \tag{5.8}$$

The first two equations lead to the learning rules for inhibitory connections and firing thresholds, once one identifies $\lambda_i \propto -\theta_i$ and $\tau_{ik} \propto -W_{ik}$; these network parameters correspond to the Lagrange multipliers of the constrained optimization problem. This reflects the fact that the role of the variable thresholds and inhibitory connections is to enforce the sparseness and non-correlation constraints in the network, which is the same as the role of the Lagrange multipliers in the Lagrange function.

I emphasize that the terms of my objective function that effectively enforce these constraints are critical for my algorithm's success. By contrast, consider the situation in which the model units had no other possibility but to maintain their fixed firing rate and lack of correlation, due to some clever parameterization of the model's state space. In that case, one could simply minimize the reconstruction error, via gradient descent, and the existence of these extra terms, or even of the analogous La-

grange multipliers, would be redundant. However, in my model, each change of the feed-forward weights (Q_{ij}) could change the neuron's firing rate, and the correlation between its activity and those of other neurons, unless something forces the network back towards the constraint surface. The variable firing thresholds and inhibitory inter-neuronal connection strengths in the model perform this function.

As discussed in the previous section, since the neuronal activities are both sparse and uncorrelated (owing to the constraints), the learning rule for the feed-forward weights Q_{ij} simplifies to one that is synaptically local (Eq. 5.5). I thus observe that the learning rules in Eq. 5.6 amount to gradient descent on a Lagrange function such that (once the network converges), its activities form optimal linear generative models of input stimuli, while maintaining sparse and decorrelated neuronal activities. Note that, while the linear generative model is optimized, given the dynamics that convert image inputs into neuronal outputs, the specific reconstruction error values might depend quite sensitively on the particular choice of dynamics.

In Fig. 5.2, I demonstrate that, once the network is trained, the activity of the SAILnet units can be linearly decoded to recover (approximately) the input stimulus.

The learning rules encourage all neurons to have the same average firing rate of μ spikes per image, which may at first appear to be at odds with the observation [Hromádka *et al.*, 2008; Baddeley *et al.*, 1997] that cortical neurons display a broad distribution of activities — firing rates vary from neuron to neuron.

However, when trained on natural images, neurons in SAILnet can actually exhibit a fairly broad range of firing rates. Moreover, the mean firing rate distribution ranges from approximately lognormal to exponential in response to natural image stimuli, depending on the mean contrast of the stimulus ensemble with which they are probed. I discuss this further in the Firing Rates section below.

I emphasize here that each of the learning rules is "synaptically" local: the information required to determine the change in the connection strength at any synaptic junction between two units is merely the activity of the pre- and post-synaptic units. The inhibitory lateral connection strengths, for example, are modified according to how many spikes arrived at the synapse, and how many times the post-synaptic unit spiked. The information required for the unit to modify its firing threshold is the unit's own firing rate. Finally, the rule for modifying the feed-forward connections requires only the pre-synaptic activity I_j, the post-synaptic activity A_i, and the present strength of that connection Q_{ij}.

5.2.3 Training SAILnet

I initialize each simulation with all inhibitory connection strengths W_{ik} set to zero, all firing thresholds θ_i set to 5, and the feed-forward weights Q_{ij} initialized with Gaussian white noise. To train the network, batches of 100 images with zero mean,

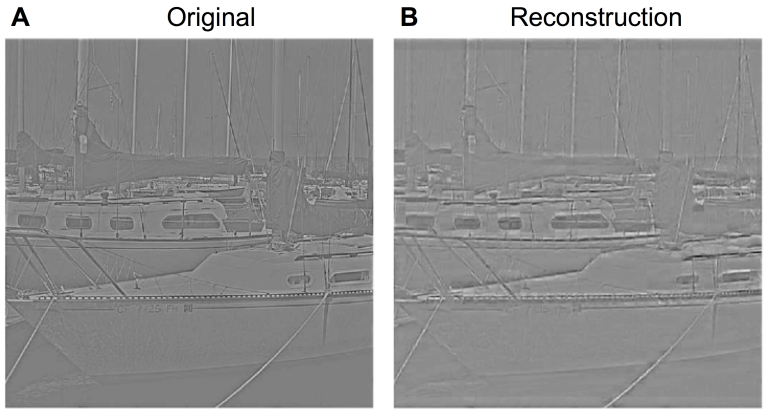

Figure 5.2: **SAILnet activity can be linearly decoded to approximately recover the input stimulus** (A) An example of a whitened image, which was not included in the training set. (B) A reconstruction of the whitened image in (A), by linear decoding of the firing rates of SAILnet neurons, which were trained on a different set of natural images. The input image was divided into non-overlapping 16×16 pixel patches, each of which was preprocessed so as to have zero-mean and unit variance of the pixel values (like the training set). Each patch was presented to SAILnet, and the number of spikes were recorded from each unit in response to each patch. A linear decoding of SAILnet activity for each patch $\overline{I_j} = \sum_i A_i Q_{ij}$ was formed by multiplying each unit's activity by that unit's RF and summing over all neurons. The preprocessing was then inverted, and the patches were tiled together to form the image in panel (B). The decoded image resembles the original, but is not identical, owing to the severe compression ratio; on average, each 16×16 input patch, which is defined by 256 continuous-valued parameters, is represented by only 75 binary spikes of activity, emitted by a small subset of the neural population. Linear decodability is a product of the SAILnet learning rules, and it is an observed feature of multiple sensory systems.

and unit standard deviation pixel values, are presented, and the number of spikes from each neuron are counted separately for each image. After each batch, the average update for the network properties is computed (following the above learning rules) over the 100-image batch. This batch-wise training lets me use matrix operations for computing the updates, which dramatically speeds up the training process. After each update, all negative values for inhibitory connections W_{ik} (which would correspond to excitatory connections) are set to zero, as in the previous work by Földiák [Földiák, 1990]. Relaxing this constraint, and allowing the recurrent weights to change sign does not affect my qualitative conclusions. In that case, some of the recurrent connections become excitatory, while the majority remain inhibitory, the RFs are qualitatively

unchanged, and the distributions of inhibitory and excitatory connection strengths are both approximately lognormal (data not shown).

The relative values of α, β and γ were chosen based on Földiák's observation that β must be much less than α or γ so that the neurons' activities remain sparse and uncorrelated, even in the face of changing feed-forward weights [Földiák, 1990].

I study the network after the properties stop changing macroscopically over time. However, as noted in the firing rates section of this paper, the network parameters continue to bounce around the final "target" state, with the size of the bounces determined by the learning rates in the network. Empirically, I find that it takes on the order of 10^7 image presentations (10^5 steps of 100 image presentations per step) for this dynamic equilibrium to be established, for the set of learning rates that I used (below). For the results presented in this paper, I let the network train for roughly 2×10^8 image presentations.

To speed up the simulation, I start the training with large values for the learning rates, and these are eventually reduced. For the last 10^4 batches of training (10^6 image presentations), the learning rates were $(\alpha, \beta, \gamma) = (0.1, 0.001, 0.01)$.

5.3 When trained on natural images, SAILnet model neurons learn the full diversity of receptive fields displayed by V1 simple cells

The RFs of 196 randomly selected units from SAILnet are shown in Fig. 5.3, as measured by their spike-triggered average activity in response to whitened natural images. These are virtually identical to the feed-forward weights of the units. This can be understood by considering the equilibrium point of the learning rule for the feed-froward weights Q_{ij} in which

$$\begin{aligned} \langle \Delta Q_{ij} \rangle &\propto \langle A_i \left(I_j - A_i Q_{ij} \right) \rangle = 0 \\ \implies \langle A_i I_j \rangle &= \left\langle A_i^2 Q_{ij} \right\rangle = \left\langle A_i^2 \right\rangle Q_{ij}, \end{aligned} \tag{5.9}$$

where the second equality occurs because the learning has converged, and thus the feed-forward weights are constant over repeated image presentations. Thus, $\langle A_i I_j \rangle / \langle A_i^2 \rangle = Q_{ij}$; the spike-triggered average (STA) stimulus is equivalent to the set of feed-forward weights, up to a multiplicative scaling factor that can be calculated from the spike train.

To facilitate a comparison between the SAILnet RFs, and those measured in macaque V1 (courtesy of D. Ringach), I fit both the SAILnet, and the macaque RFs

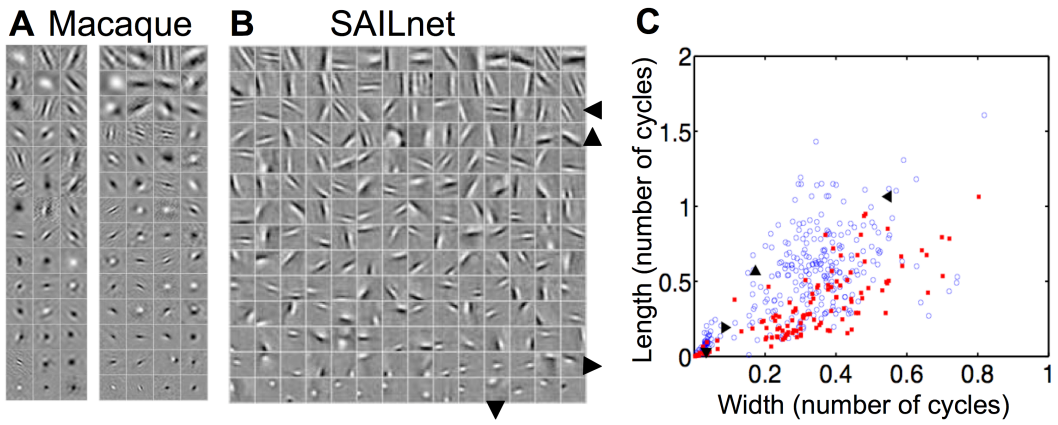

Figure 5.3: **SAILnet learns receptive fields (RFs) with the same diversity of shapes as those of simple cells in macaque primary visual cortex (V1)** (A) 98 randomly selected receptive fields recorded from simple cells in macaque monkey V1 (courtesy of D. Ringach). Each square in the grid represents one neuronal RF. The sizes of these RFs, and their positions within the windows, have no meaning in comparison to the SAILnet data. The data to the right of the break line have an angular scale (degrees of visual angle spanned horizontally by the displayed RF window) of 0.94^o, whereas those to the left of it span 1.88^o. (B) RFs of 196 randomly selected model neurons from a 1536-unit SAILnet trained on patches drawn from whitened natural images. The gray value in all squares represents zero, whereas the lighter pixels correspond to positive values, and the darker pixels correspond to negative values. All RFs are sorted by a size parameter, determined by a Gabor function best fit to the RF. The SAILnet model RFs show the same diversity of shapes as do the RFs of simple cells in macaque monkey V1 (A); both the model units and the population of recorded V1 neurons consist of small unoriented features, oriented Gabor-like wavelets containing multiple subfields, and elongated edge-detectors. (C) I fit the SAILnet and macaque RFs to Gabor functions, in order to quantify their shapes. Shown are the dimensionless *width* and *length* parameters ($\sigma_x \times f$ and $\sigma_y \times f$, respectively) of the 299 SAILnet RFs and 116 (out of 250 RFs in the dataset) macaque RFs for which the Gabor fitting routine converged. These parameters represent the size of the Gaussian envelope in either direction, in terms of the number of cycles of the sinusoid. The SAILnet data (open blue circles) span the space of the macaque data (solid red squares) from the Gabor fitting analysis; SAILnet is accounting for all of the observed RF shapes. I highlight four SAILnet RFs with distinct shapes, which are identified by the large triangular symbols that are also displayed next to the corresponding RFs in panel (B).

to Gabor functions. As in the SSC study of Rehn and Sommer, only those RFs that could be sensibly described by a Gabor function were included in Fig. 5.3; for example, I excluded RFs with substantial support along the square boundary, suggesting that the RF is only partly visible. In [Zylberberg *et al.*, 2011], I discuss the Gabor fitting routine and the quality control measures I used to define and identify meaningful fits.

The SAILnet model RFs show the same diversity of shapes observed in macaque V1, and in the non-local SSC model [Rehn and Sommer, 2007]. They consist of three qualitatively distinct classes of neuronal RFs: small unoriented features, localized and oriented Gabor-like filters, and elongated edge-detectors. The SAILnet learning rules approximately minimize the same cost function as the SSC model, albeit with constraints as opposed to unconstrained optimization, which explains how it is possible for SAILnet to learn similar RFs using only local rules. Furthermore, in my model, the number of co-active units is small, owing to the low average lifetime neuronal firing rates, and the fact that spikes can only be emitted in integer numbers. This feature is similar to the L_0-norm minimization used in the SSC model of Rehn and Sommer and the LCA model of Rozell and colleagues [Rozell *et al.*, 2008].

When it was first published, this was the first demonstration that a network of spiking neurons using only synaptically local plasticity rules applied to natural images can account for the observed diversity of V1 simple cell RF shapes [Zylberberg *et al.*, 2011].

The specific shapes of the SAILnet receptive fields depend on both the mean firing rate of the units, μ, as well as the degree of overcompleteness of the network. A rough attempt to optimize these values to best fit the macaque data was performed manually, but a full search over the space of {overcompleteness, μ} would be too time consuming. As a result, it is likely that even tighter quantitative agreement between the SAILnet, and macaque receptive fields is possible: while the SAILnet RFs shown in Fig. 5.3 show the full set of shapes displayed by the macaque RFs, SAILnet also shows many RFs that are longer and narrower than the macaque RFs. In that sense, the observed SAILnet RFs are a superset of the macaque ones.

5.4 SAILnet units can exhibit a broad distribution of mean firing rates in response to natural images

The SAILnet learning rules encourage every unit to have the same target value, μ, for its average firing rate, which might appear to be inconsistent with observations [Hromádka *et al.*, 2008; Baddeley *et al.*, 1997; Abeles *et al.*, 1990] that cortical neurons exhibit a broad distribution of mean firing rates. However, I will demonstrate herein that SAILnet, too, can display a wide range of mean rates.

To determine the distribution of mean firing rates across the population of model

neurons in the network, I studied the fully trained network. The measurement was then performed with all learning rates set to zero, so that I was probing the properties of the network at one fixed set of learned parameter values, rather than observing changes in network properties over time.

I then counted the number of spikes per image from each unit to estimate each neuron's average firing rate, as it might be measured in a physiology experiment. The distribution of these mean firing rates is fairly broad and well-described by a lognormal distribution (Fig. 5.4). This distribution is strongly non-monotonic, clearly indicating that it is poorly fit by an exponential function.

Subsequently, I probed the same network (still with the learning turned off, so that the network parameters were identical in both cases) with $50,000$ low-contrast images consisting of patches from our training ensemble with all pixel values multiplied by $1/3$. I found that the firing rate distribution was markedly different than what I found when the network was probed with higher-contrast stimuli. In particular, it became a monotonic decreasing function that was similarly well-described by either a lognormal or an exponential function.

From the dynamics of the leaky integrate-and-fire units, it is clear that the low contrast stimuli with reduced pixel values will cause the units to charge up more slowly and subsequently to spike less in the allotted time the network is given to view each image. Consequently, the firing rate distribution gets shifted towards lower firing rates. However, negative firing rates are impossible, so in addition to being shifted, the low-firing-rate tail of the distribution is effectively truncated. Note that truncating the lognormal distribution anywhere to the right of the peak results in a distribution that looks qualitatively similar to an exponential.

Mean firing rates in primary auditory cortex (A1) have been reported to obey a lognormal distribution [Hromádka *et al.*, 2008] , whether spontaneous or stimulus-evoked in both awake and anesthetized animals. However, exponentially distributed spontaneous mean firing rates have also been reported in awake rat A1 [Gaese and Ostwald, 2003]. Although several groups have measured the distribution of firing rates over time for individual neurons [Baddeley *et al.*, 1997; Abeles *et al.*, 1990], I am unaware of a published claim regarding the distribution of mean firing rates in visual cortex.

Recall that the SAILnet learning rules encourage the neurons to all have the same average firing rate. This fact may be puzzling at first given the spread in mean firing rates apparent in the distributions shown in Fig. 5.4. There are two main effects to consider when making sense of this: finite measurement time, and non-zero step-sizes for plasticity.

The first effect relates to the fact that there is intrinsic randomness in the measurement process — which randomly selected image patches happen to fall in the ensemble of probe stimuli — so that the measured distribution tends to be broader

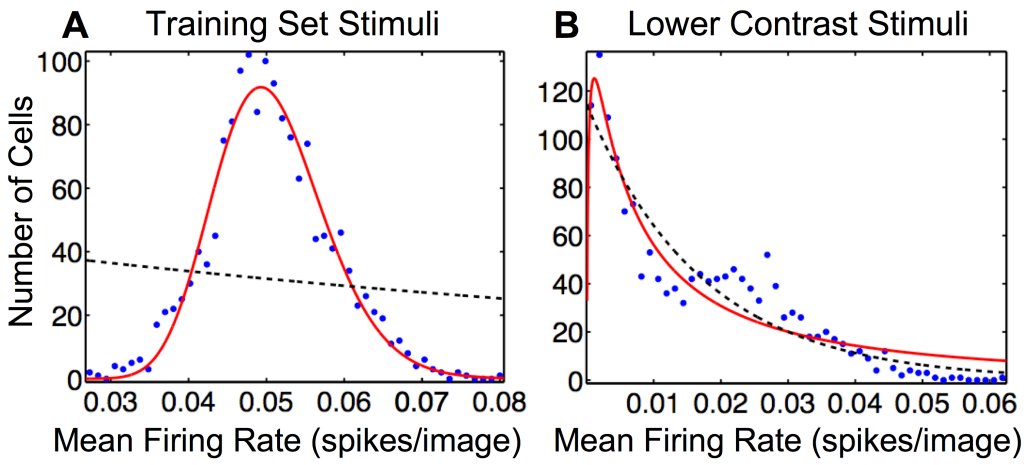

Figure 5.4: **Units in SAILnet exhibit a broad range of mean firing rates, which can be lognormally or exponentially distributed depending on the choice of probe stimuli.** (A) Frequency histogram of firing rates averaged over $50,000$ image patches drawn from the training ensemble for each of the 1536 units of a SAILnet trained on whitened natural images. All learning rates were set to zero during probe stimulus presentation. A wide range of mean rates was observed, but as expected, the distribution is peaked near $\mu = 0.05$ spikes per image, the target mean firing rate of the neurons. The paucity of units with near-zero firing rates suggests that this distribution is closer to lognormal than exponential. Accordingly, the lognormal least-squares (solid red curve) fit accounts for $R^2 = 96$ % of the variance in the data, whereas the exponential fit (black dashed curve) accounts for only 2 %. (B) In response to low contrast stimuli, the firing rate distribution across the units (every unit fired at least once) in the same network as in panel (A) was similarly well fit by either an exponential (dashed black curve; accounting for $R^2 = 88$ % of the variance in the data) or a lognormal function (solid red curve; accounting for 90 % of the variance). The low-contrast stimulus ensemble used to probe the network consisted of images drawn from the training set, with all pixel values reduced by a factor of three.

than the "true" underlying distribution of the system. To check that this effect is not responsible for the broad distribution in firing rates, I computed the variance in the measured firing rate distribution after different numbers of images were presented to the network. The variance decreased until it reached an asymptotic value after approximately $25,000$–$30,000$ image presentations (data not shown). Thus, the $50,000$ image sample size in my experiment is large enough to see the true distribution; finite sample-size effects do not affect the distributions that I observed.

The other, more interesting, effect that gives rise to a broad distribution of firing rates is related to learning. While the network is being trained, the feed-forward weights, inhibitory lateral connections, and firing thresholds get modified in discrete jumps, after every image presentation (or every batch of images). Since those jumps are of a non-zero size – with their magnitudes determined by the learning rates α, β, and γ – there will be times when the firing threshold gets pushed below the specific value that would lead to the unit having exactly the target firing rate, and the unit will thus spike more than the target rate. Similarly, some jumps will push the threshold above that specific value, and the unit will fire less than the target amount. Even after learning has converged, and the parameters are no longer changing *on average* in response to additional image presentations, the network parameters are still bouncing around their average (optimal) values; any image presentation that makes a neuron spike more than the target amount results in an increased firing threshold, while any image that makes the neuron fire less than the target amount leads to a decreased firing threshold. Recent results suggest that the sizes of these updates (jumps) are quite large for real neurons [Clopath *et al.*, 2010]. Interestingly, this indicates that the observed broad distributions in firing rate [Hromádka *et al.*, 2008] do not rule out the possibility that homeostatic mechanisms are driving each neuron to have the same average firing rate.

Reducing the SAILnet learning rates α, β and γ does reduce the variance of the firing rate distributions, but the qualitative conclusions — non-monotonic, approximately lognormal firing rate distribution in response to images from the training set, and monotonic, exponential/lognormal distribution in response to low contrast images — are unchanged when I use different learning rates for the network (data not shown).

5.5 Pairs of SAILnet units have small firing rate correlations.

Recent experimental work has shown that neurons in visual cortex tend to have small correlations between their firing rates [Ecker *et al.*, 2010; Renart *et al.*, 2010]. In order to facilitate a comparison between the SAILnet model, and the physiological observations, I measured the (Pearson's) linear correlation coefficients between spike counts of SAILnet units, in response to an ensemble 30,000 natural images. These correlations (Fig. 5.5) tend to be near zero, as is observed experimentally [Ecker *et al.*, 2010], while the experimental data show a larger variance in the distribution of correlation coefficients than I observe with SAILnet. Like the firing rate distribution (discussed above), the distribution of correlation coefficients is affected by the update sizes (learning rates) in the simulation, with larger update sizes leading to a larger

variance of the measured distribution.

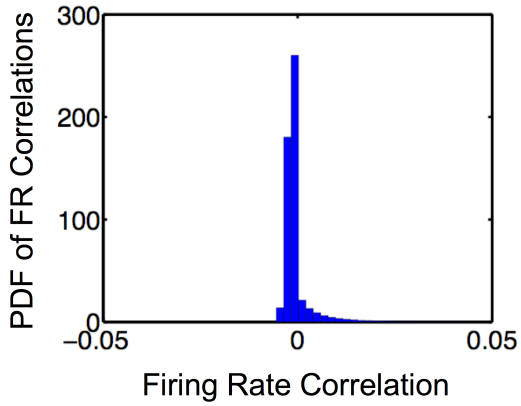

Figure 5.5: **Pairs of SAILnet units have small firing rate correlations.** The probability distribution function (PDF) of the Pearson's linear correlation coefficients between the spike-counts of pairs of SAILnet neurons responding to an ensemble of 30,000 natural images is sharply peaked near zero.

In Fig. 5.5, the distribution appears truncated on the left. This effect arises because there is a lower bound on the correlation between the neuronal firing rates that arises when the two neurons are *never* co-active. The low mean firing rate of $\mu = 0.05$ spikes per neuron per image used in the simulation means that this bound is not too far below zero.

5.6 Connectivity learned by SAILnet allows for further experimental tests of the model

Several previous studies of sparse coding models have focused on the receptive fields learned by adaptation to naturalistic inputs [Olshausen and Field, 1996; Rehn and Sommer, 2007; Földiák, 1990; Falconbridge *et al.*, 2006; Perrinet *et al.*, 2004; Olshausen *et al.*, 2009] , but I am aware of only one published study [Garrigues and Olshausen, 2008] that investigated the connectivity in sparse coding models, albeit with a model that lacked biological realism. One previous study [Koulakov *et al.*, 2009] investigated synaptic mechanisms that could give rise to the measured distribution of connection strengths, but this work was not performed in the context of a sensory coding model. No prior work has studied the connectivity learned in a biophysically well-motivated sensory coding network, which would provide additional testable predictions for physiology experiments.

Fig. 5.6 shows the distribution of non-zero connection strengths (non-zero elements of the matrix W_{ik}) learned by a 1536-unit SAILnet with $\mu = 0.05$ trained on 16×16 pixel patches drawn from whitened natural images. When trained on natural images, SAILnet learns an approximately lognormal distribution of inhibitory connection strengths; a Gaussian best fit to the histogram of the logarithms of the connection strengths accounts for 98% of the variance in the data.

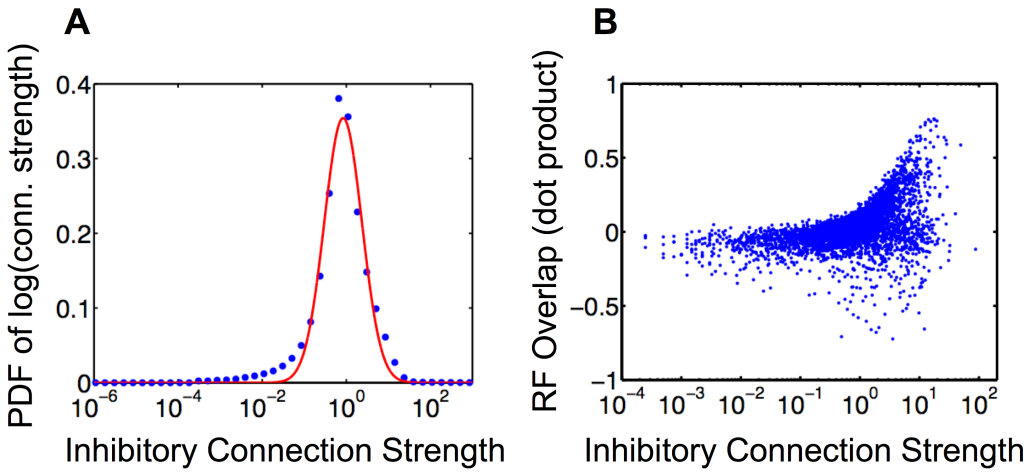

Figure 5.6: **Connectivity learned by SAILnet allows for further experimental tests of the model.** (A) Probability Density Function (PDF) of the logarithms of the inhibitory connection strengths (non-zero elements of the matrix W_{ik}) learned by a 1536 unit SAILnet trained on 16×16 pixel patches drawn from whitened natural images. The measured values (blue points) are well-described by a Gaussian distribution (solid line), which accounts for $R^2 = 98\ \%$ of the variance in the dataset. This indicates that the data are approximately lognormally distributed. Note that there are some systematic deviations between the Gaussian best fit and the true distribution, particularly on the low-connection strength tail, similar to what has been observed for excitatory connections within V1. This plot was created using the binning procedure of Hromádka and colleagues. The histogram was normalized to have unit area under the curve. (B) The strengths of the inhibitory connections between pairs of cells are correlated with the overlap between those cells' receptive fields: cells with significantly overlapping RFs tend to have strong mutual inhibition. Data shown in panel (B) are for 5,000 randomly selected pairs of cells. Pairs of cells with significantly negatively overlapping RFs tend not to have inhibitory connections between them, hence the apparent asymmetry in the RF overlap distribution obtained by marginalizing over connection strengths in panel (B).

Despite this close agreement, SAILnet shows some systematic deviations from the lognormal fit, especially on the low-connection-strength tail of the distribution. Interestingly, the experimental data [Song. *et al.*, 2005] show an approximately lognormal distribution of excitatory connection strengths, with similar systematic deviations (Fig. 5b of [Song. *et al.*, 2005]). By contrast, prior theoretical work [Koulakov *et al.*, 2009; Song. *et al.*, 2005] has employed learning rules tailored to create exactly lognormal connection strength distributions, and thus show no such deviations. Note also that neither of these previous studies addressed the issue of how neurons might represent sensory inputs, nor how they might learn those representations.

Whereas the experimental data of Song *et al.* show a roughly lognormal distribution in the strengths of excitatory connections between V1 neurons, the SAILnet model makes predictions about the strengths of *inhibitory* connections in V1. The 1 ms time window for measuring post-synaptic potentials in the experiment of Song *et al.* ensured that they measured only direct synaptic connections. However, suppressive interactions between excitatory neurons in cortex are mediated by inhibitory interneurons. Consequently, the inhibitory interactions between pairs of excitatory neurons in V1 must involve two or more synaptic connections between the cells. Thus, my model predicts that the inhibitory functional connections between excitatory simple cells in V1, like the excitatory connections measured by Song *et al.*, should follow an approximately lognormal distribution (Fig. 5.6), but it does not specify the extent to which this is achieved through variations in strength among dendritic or axonal synaptic connections of V1 inhibitory interneurons. One recent theoretical study [Clopath *et al.*, 2010] has uncovered some interesting relationships between coding schemes and connectivity in cortex, but it did not make any statements about the anticipated distribution of inhibitory connections.

Interestingly, there is a clear correlation between the strengths of the inhibitory connection between pairs of SAILnet neurons, and the overlap (measured by vector dot product) between their receptive fields: neurons with significantly overlapping receptive fields tend to have strong inhibitory connections between them (Fig. 5.6). This correlation is expected because cells with similar RFs receive much common feed-forward input. Thus, in order to keep their activities uncorrelated, significant mutual inhibition is required. This same feature was assumed by the LCA algorithm of Rozell and colleagues, but is naturally learned by SAILnet, in response to natural stimuli.

The connectivity predictions are amenable to direct experimental testing, although that testing may be challenging, owing to the difficulty of measuring functional connectivity mediated by two or more synaptic connections between pairs of V1 excitatory simple cells.

5.7 Discussion

The present work demonstrates that synaptically local plasticity rules can be used to learn a sparse code for natural images that accounts for the diverse shapes of V1 simple cell receptive fields. The SAILnet model uses purely synaptically local learning rules — connection strengths are updated based only on the number of spikes arriving at the synapse and the number of spikes generated by the post-synaptic cell. By contrast, the local competition algorithm (LCA) of Rozell and colleagues assumes that $W_{ik} = \sum_j Q_{ij} Q_{kj}$, so that the strength of the inhibitory connection between two neurons is equal to the overlap (*i.e.*, vector dot product) between their receptive fields [Rozell *et al.*, 2008]. This non-local rule requires that individual inhibitory synapses must somehow keep track of the changes in the receptive fields of many neurons throughout the network in order to update their strengths. Moreover, the LCA network does not contain spiking units, even though cortical neurons are known to communicate via discrete, indistinguishable, spikes of activity [Dayan and Abbott, 2001].

Similarly, the units in the networks of [Falconbridge *et al.*, 2006] and [Földiák, 1990] communicate via continuous-valued functions of time. Although these two models do use synaptically local plasticity rules, neither of these groups demonstrated that such local plasticity rules are sufficient to explain the diversity of simple cell RF shapes observed in V1, nor have they determined the objective function optimized by their learning rules.

Independent of the present work [Shapero *et al.*, 2011] have recently implemented a spiking version of LCA [Rozell *et al.*, 2008] that uses integrate-and-fire units. However, that work does not address the issue of how to train such a network using synaptically local plasticity rules.

Some groups have used spiking units to perform image coding [Perrinet *et al.*, 2004; Perrinet, 2010; VanRullen and Thorpe, 2002; Masquelier and Thorpe, 2007; Delorme *et al.*, 2000], but those studies did not address the question of whether synaptically local plasticity rules can account for the observed diversity of V1 RF shapes. Interestingly, it has been demonstrated [Delorme *et al.*, 2000] that orientation selectivity can arise from spike timing dependent plasticity rules applied to natural scenes. Previous work [Perrinet, 2010] has also explored the addition of homeostatic mechanisms to sparse coding algorithms and found it to improve the rate at which learning converges and to qualitatively affect the shapes of the learned RFs; homeostasis is enforced in the SAILnet model via modifiable firing thresholds.

Finally, one previous group [Savin *et al.*, 2010] has demonstrated that independent component analysis (ICA) can be implemented with spiking neurons and local plasticity rules. That work did not, however, account for the diverse shapes of V1 receptive fields, although they did also demonstrate that homeostasis (a mean firing

rate constraint) was critical to the learning process.

The SAILnet model attempts to be biophysically realistic, but it is not a perfect model of visual cortex in all of its details. In particular, like many previous models [Olshausen and Field, 1996; Rehn and Sommer, 2007; Perrinet *et al.*, 2004; Perrinet, 2010; Olshausen *et al.*, 2009], the network alternates between brief periods of inference (the representation of the input by a specific population activity pattern in the network) and learning (the modification of synaptic strengths), which may not be realistic. Indeed, it is unclear how cortical neurons would "know" when the inference period is over and when the learning period should begin, though it is interesting to note that these iterations could be tied to the onset of saccades, given the $5\tau_{RC} \approx 100\ ms$ inference period between "learning" stages in our model.

As in previous models, the inputs to SAILnet, I_j, are continuous-valued, whereas the actual inputs from the lateral geniculate nucleus to primary visual cortex (V1) are spiking. As mentioned above, suppressive interactions between pairs of units in the model are mediated by direct, one-way, inhibitory synaptic connections between units, rather than being mediated by a distinct population of inhibitory interneurons. I did not include the effects of spike-timing dependent plasticity [Dan and Poo, 2006], although this has been shown to have interesting theoretical implications for cortex [Clopath *et al.*, 2010] in general and for image coding in particular [Masquelier and Thorpe, 2007; Delorme *et al.*, 2000]. An interesting avenue for future research is to develop models that incorporate spike timing dependent learning rules, applied to time-varying image stimuli such as natural movies.

Finally, the neurons in my model have no intrinsic noise in their activities, although that noise may, in practice, be small [Mainen and Sejnowski, 1995].

Interestingly, since the SAILnet model neurons require a finite amount of time to update their internal variables $u_i(t)$, there is a hysteresis effect if one presents the network with time-varying image stimuli — the content of previous frames affects how the network processes and represents the current frame. Even if the features in a movie change slowly, the optimal representation of one frame can be very different from the optimal representation of the next frame in many image coding models, so this hysteresis effect can provide stability to the image representation compared to other models such as ICA [Bell and Sejnowski, 1997; Hyvärinen and Hoyer, 2001] or Olshausen and Field's Sparsenet [Olshausen and Field, 1996]. This effect has previously been studied by Rozell and colleagues [Rozell *et al.*, 2008], encouraging future efforts to apply SAILnet to dynamic stimuli.

Though it is highly simplified, the SAILnet model does captures many qualitative features of V1, such as inhibitory lateral connections [Haider *et al.*, 2010], largely uncorrelated neuronal activities [Ecker *et al.*, 2010; Renart *et al.*, 2010], sparse neuronal activity [Haider *et al.*, 2010; Vinje and Gallant, 2002; Vinje and Gallant, 2000; Hromádka *et al.*, 2008; Baddeley *et al.*, 1997], a greater number of cortical neurons

than input neurons (overcomplete representation), synaptically local learning rules, and spiking neurons. Importantly, this model allows me to make several falsifiable experimental predictions about interneuronal connectivity and population activity in cortex. I hope that these predictions will help uncover the coding principles at work in the visual cortex.

5.7.1 Information theoretic discussion of SAILnet

Unlike ICA and sparse coding, which both explicitly model the probability distribution function over natural images – and thus have statistical independence as one of their objectives – SAILnet instead seeks the more modest goal of pair-wise decorrelation. As I have previously discussed, maximum information-theoretic efficiency comes with statistically independent units, yet decorrelation is no guarantee of full independence. In that sense, SAILnet is not as sophisticated as either ICA, or the sparse coding models.

At the same time, one of the main arguments in favor of sparse coding is that of metabolic efficiency: that, because brains are energetically expensive, evolution should favor designs that minimize this cost. In that regard, SAILnet fares quite well, inasmuch as its objective is to form the best linear generative model it can, for a given level of energetic cost (defined by the firing rate constraint). Clearly, this metric is closely related to the information-theoretic one, so SAILnet may not be quite as far behind the canonical sparse coding models as it may appear at first blush.

Regardless of these efficiency arguments, my SAILnet model is sufficient to demonstrate that synaptically local plasticity rules can suffice to form an optimal linear generative code for natural images, and that such a model can account for the full diversity of shapes exhibited by V1 receptive fields.

The Gabor fitting routine used to generate the scatter plot in Fig. 5.3 was provided by Jason Murphy, and the macaque V1 receptive fields in Fig. 5.3 are courtesy of Dario Ringach.

Chapter 6

Why does sparseness in ferret V1 decrease during development?

6.1 Decreasing sparseness during development is in conflict with the canonical sparse coding models

As I have shown in previous chapters, sparseness is an appealing theoretical concept, which leads to efficient cortical representations for natural images, and one for which there is a reasonable amount of experimental support. At the same time, genetic and developmental evidence suggest that visual cortical physiology is a product of experience, and thus the theoretical models should be compatible with both the physiology of the mature animal, and the developmental trends leading up to that state. Finally, since it is unclear what absolute sparseness level is required to support the sparse coding hypothesis, one might instead choose to study its evolution during the development of the animal: how does sparseness change as a result of adaptation and experience? For typical sparse coding models (*e.g.*, [Olshausen and Field, 1996; Rehn and Sommer, 2007] and related models), neurons gradually learn features that allow for a sparser encoding of the stimuli, so the sparseness invariably increases over time (Fig. 6.1). Physiology experiments, however, show something very different in the developing visual cortex. Recently, Berkes and colleagues measured multi-unit V1 activity in awake young ferrets viewing natural movies, and found that, as the animals matured, their stimulus-driven V1 activity became *less* sparse [Berkes *et al.*, 2009] (Fig. 6.1), in stark contrast with the computational models [Rehn and Sommer, 2007; Olshausen and Field, 1996; Bell and Sejnowski, 1997]. Conversely, it has also been observed that spontaneous slow-wave activity (recorded in the absence of a visual stimulus) in the anesthetized mouse visual cortex becomes sparser immediately after eye-opening [Rochefort *et al.*, 2009]. While these experimental findings may seem

contradictory, spontaneous activity is not as easily related to the sparse coding hypothesis, which does not have much to say about activity in the absence of sensory input.

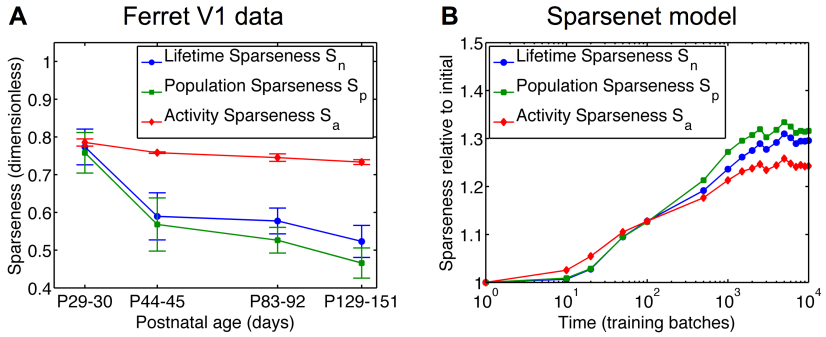

Figure 6.1: **V1 developmental data pose a challenge to canonical sparse coding models.** Multi-unit activity in primary visual cortex (V1) of awake young ferrets watching natural movies show decreasing sparseness over time (**A**) . The sparseness metrics shown in this figure are defined in section 6.2, and the data are courtesy of Pietro Berkes. For comparison, the canonical Sparsenet model of Olshausen and Field shows the opposite trend (**B**), for which sparseness increases as a result of experience with viewing natural scenes. Both plots have logarithmic horizontal axes.

One natural explanation for the ferret developmental data might involve homeostasis, in which the cortex seeks to maintain some fixed target level of activity. If the developmental initial conditions resulted in neuronal activities that were below the target activity level, regulation might lead to decreasing sparseness.

Indeed, homeostatic activity regulation has been widely reported in visual cortical neurons [Mrsic-Flogel *et al.*, 2007; Hendry and Jones, 1988; Marder and Goaillard, 2006; Desai *et al.*, 1999; Turrigiano *et al.*, 1998; Turrigiano and Nelson, 2000; Burrone and Murthy, 2003; Davis and Bezprozvanny, 2001]. Specifically, when cultured visual cortical pyramidal cells have their firing rates pharmacologically altered — by blocking GABA-mediated inhibition, for example — adaptation returns the cells to their original firing rates, without the removal of the pharmacological agents [Turrigiano *et al.*, 1998]. Moreover, when deprived of stimulation, such cultured cells become more excitable [Desai *et al.*, 1999]. Similar observations have been made *in vivo* [Mrsic-Flogel *et al.*, 2007]. One might then wonder whether a computational model in which neuronal activity is regulated by homeostasis might be able to account for both the receptive field shapes of V1 simple cells, which are correctly predicted by the sparseness-maximizing models [Olshausen and Field, 1996;

Rehn and Sommer, 2007; Bell and Sejnowski, 1997], and the decreasing sparseness observed during V1 development [Berkes *et al.*, 2009], which is in conflict with sparseness-maximizing models.

In the preceding chapter, I introduced a computational model for visual cortex that could successfully reproduce the full diversity of RFs exhibited by V1 simple cells while using spiking neurons and synaptically local plasticity rules [Zylberberg *et al.*, 2011]. The model is computationally well-understood, in that learning in the model can be shown to maximize the fidelity with which the neural activities represent input stimuli subject to constraints on the mean firing rates and pair-wise neuronal correlations. In order for these synaptically local plasticity rules to optimize the cooperative representation of input stimuli by the neural population, the activities must be sparse and uncorrelated throughout learning [Zylberberg *et al.*, 2011], hence the constraints. These constraints are enforced in the model by homeostatic mechanisms that regulate the neural activity levels via variable firing thresholds — equivalent to synaptic scaling [Abbott and Nelson, 2000] or modifiable intrinsic excitability [Turrigiano, 2011] for the leaky integrate-and-fire model neurons — and modifiable inhibitory lateral connections that lead to uncorrelated neuronal discharges. In order for the neurons in the homeostatic SAILnet model to learn RFs that agree with physiology experiments, this target activity level must be low, so the neuronal responses are sparse, despite the fact that they were not made so by an optimization mechanism that necessarily results in increasing sparseness during learning. Specifically, if the initial conditions of that model result in neuronal activities that are below the target activity level, learning makes the responses less sparse over time, in agreement with the ferret V1 data [Berkes *et al.*, 2009], whereas initial conditions producing activity exceeding the target level result in increased sparseness over time.

Thus, since the homeostatic SAILnet model can account for the decreased sparseness during development, in addition to the specific diversity of RFs reported in V1, it might provide a more parsimonious explanation for V1 physiology than previous models involving active maximization of sparseness. The importance of sparseness and decorrelation in allowing synaptically local plasticity rules to optimize cooperative representations in the model suggests that homeostasis may play a surprisingly deep role in the learning of sensory representations.

In the remainder of this chapter, I will provide evidence to support these claims.

6.2 SAILnet single- and multi-unit activity can become less sparse during receptive field formation, in agreement with V1 development

To study the change in sparseness over time, I ran SAILnet simulations, starting with randomized feed-forward weights, recurrent connection strengths, and firing thresholds that were initialized with Gaussian-distributed white noise: I give values for all of these parameters in the Methods section at the end of this chapter. At different times during the adaptation process, I recorded the simulated neuronal activity in response to randomly selected batches of natural images. Following a recent experimental study [Berkes *et al.*, 2009], I computed from these network activities three sparseness measures. First, the "activity sparseness" (S_a), which is the fraction of units inactive in response to any given stimulus:

$$S_a = 1 - n_a/N, \tag{6.1}$$

where n_a is the number of neurons whose activities rose above some threshold number of spikes in response to the stimulus, and N is the total number of neurons for which data were recorded. I set the threshold to 6 spikes for the single-unit sparseness data shown herein, and to 8 spikes for the "multi-unit" measurement, discussed below. I performed this measurement by averaging over 500 different input stimuli. The activity sparseness S_a is very similar to the L_0 norm of the neuronal activities.

In addition, I recorded two other sparseness measures, originally due to Treves & Rolls [Treves and Rolls, 1991] (TR), and subsequently modified by Vinje & Gallant [Vinje and Gallant, 2000]. First, consider what I will call the "TR population sparseness" measure, S_p,

$$S_p = \left[1 - \frac{\left(\sum_{i=1}^{N} |A_i|/N\right)^2}{\sum_{i=1}^{N} |A_i|^2/N}\right] \times \left(1 - \frac{1}{N}\right)^{-1}, \tag{6.2}$$

where A_i is the activity of neuron i. Note that S_p is assessed in response to a single image, although for the current purposes, I will average this measure over 500 different image stimuli, to infer the average TR population sparseness. Similarly, I will define the "TR lifetime" sparseness of a single neuron, S_n, the same way (Eq. 6.2), but with the replacement that A_i represents the neuron's activity in response to a given image i, and N will be the number of different image stimuli (500) for which activities are recorded. Similar to the TR population sparseness S_p, I will average these values over the entire population for our measurement.

In Fig. 6.2, I show the receptive fields of SAILnet neurons both before and after the network is trained with natural scenes. I also show the evolution of the sparseness measures during that training process. Contrary to the more common sparseness-maximizing models (*e.g.* [Olshausen and Field, 1996; Rehn and Sommer, 2007; Bell and Sejnowski, 1997] (see Fig. 6.1), our SAILnet model can display *decreased* sparseness by all three measures while it is learning localized and oriented receptive fields. The time course of the sparseness measures depends on the learning rates (adaptation step sizes), with smaller learning rates leading to slower changes in sparseness measures, as expected (data not shown).

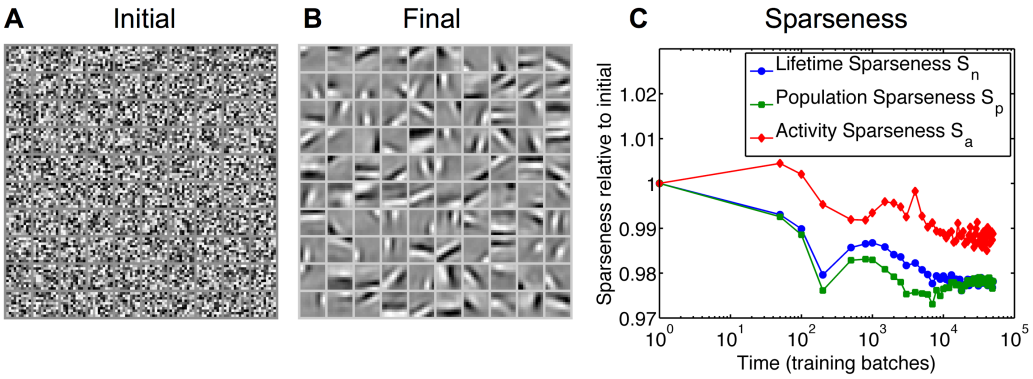

Figure 6.2: **SAILnet single-unit activity can become less sparse during receptive field formation.** A SAILnet simulation was performed in which the RFs, firing thresholds, and recurrent connection strengths were initialized with random numbers (see Methods section for details). The RFs of the 100 neurons are shown in panel **A**. Each box on the grid shows the RF of one neuron, with white corresponding to positive pixel values, and black corresponding to negative ones. After training with natural images, these same SAILnet neurons have oriented, localized RFs (**B**). All three sparseness measures decrease during the training period (**C**).

Note that I cannot quantitatively compare these model predictions (Fig. 6.2) directly with the ferret V1 developmental data reported by Berkes and colleagues because the physiology data are for *multi-unit* activity [Berkes *et al.*, 2009], whereas the results in Fig. 6.2 depict *single-unit* activity.

Multi-unit activity is recorded in extracellular physiology experiments by high-pass filtering the voltage trace, and thresholding to obtain spikes. Whereas one would then perform "spike sorting" to assign the spikes to individual neurons in order to obtain single-unit data, for multi-unit data, one simply considers the aggregate of all of the spikes observed. Multi-unit data is easier to obtain than single-unit data, does

not fall prey to systematic errors from spike sorting, and reflects the summation of the activity of many different neurons.

In order to make a more meaningful comparison between the model and experiment, I mimicked a multi-unit activity measurement by randomly grouping together sets of 4 SAILnet neurons, whose activities are then summed to form a multi-unit response. I then repeated the SAILnet simulation, and the sparseness measurements, using these group activities in place of the single-unit activities for the sparseness measurements. This procedure yielded results (Fig. 6.3) that are qualitatively similar to the single-unit activity shown in Fig. 6.2, but they show a larger change in sparseness values that is more consistent with the ferret data. A direct quantitative comparison between the multi-unit model data and the ferret data is difficult because it is not clear how best to estimate the relevant number of neurons to group together, or even whether all groupings should have the same number of neurons. However, it is apparent that SAILnet learning is in qualitative agreement with the ferret V1 development data for the right choice of initial conditions.

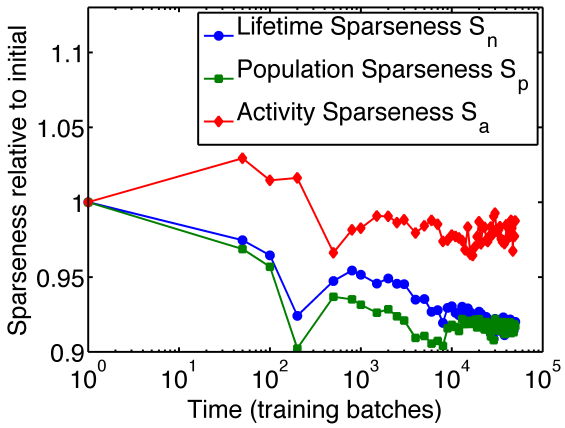

Figure 6.3: **SAILnet multi-unit activity can become less sparse during receptive field formation, in qualitative agreement with the ferret V1 developmental data.** I repeated the simulation described in Fig. 6.2, but randomly grouped together sets of 4 SAILnet neurons whose activities were summed to yield multi-unit activity for the sparseness measurements. These multi-unit activities show the same trend as the single-unit responses shown in Fig. 6.2. Specifically, all three sparseness measures decrease over time, in qualitative agreement with multi-unit activity recorded during development in ferret V1 (Fig. 6.1).

6.3 For some initial conditions, SAILnet single-unit activity becomes more sparse during receptive field formation

I find that SAILnet does not *require* sparseness to decrease over time, rather, it is *compatible* with decreasing sparseness. To demonstrate this point, I repeated the SAILnet simulations, and sparseness measurements, with different initial conditions: the feed-forward weights and recurrent connection strengths were randomized with the mean and variance of the initial recurrent strength distribution being smaller than for the results shown in Figs. 6.2 and 6.3, while firing thresholds had the same value for all neurons. For these initial conditions, but with all other model parameters including the learning rates and target neural firing rate the same as the simulation discussed above (and in Fig. 6.2), I observe an increase in sparseness during training (Fig. 6.4), similar to what one finds with Sparsenet [Olshausen and Field, 1996] (Fig. 6.1), and related models.

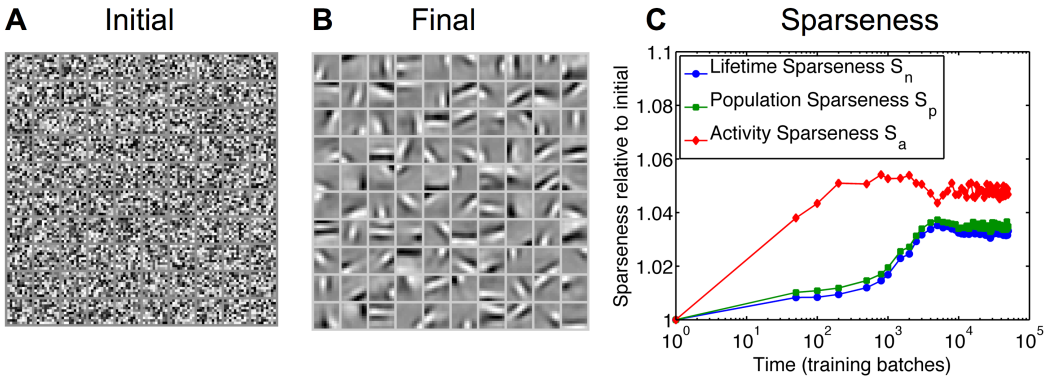

Figure 6.4: **For some initial conditions, SAILnet single unit sparseness measures increase during training.** A SAILnet simulation was performed in which the RFs were initially randomized, and the recurrent connection strengths were initialized with random numbers that were much smaller than for the simulation described in Fig. 6.2 (see Methods section for details). The RFs of the 100 neurons are shown in panel **A**. As in Fig. 6.2, each box on the grid depicts the RF of one neuron, with lighter tones corresponding to positive pixel values, and darker tones corresponding to negative values. After training with natural images, these same SAILnet neurons have oriented, localized RFs (**B**). All three sparseness measures increase during the training period (**C**). Aside from the initial conditions, the network used to generate these data was identical to the one from Fig. 6.2.

6.4 The diversity of receptive field shapes depends on the homeostatic set point of unit activity in SAILnet

When trained on natural scenes, SAILnet learns the same specific diversity of receptive fields as exhibited by macaque V1 [Zylberberg *et al.*, 2011] provided that the homeostatically controlled target firing rate for single units is set to the appropriate level. As I have previously discussed, the optimal target firing (for which the RFs give the best agreement) rate depends non-trivially on the degree of overcompleteness of the representation — the number of neurons compared to the number of pixels in the input image frames. The specific shapes and sizes of the RFs learned by SAILnet also depend on the value of the target firing rate, however, and when the firing rates are set too low, the dictionary of learned RF shapes is qualitatively different from what is seen in V1 (Fig. 6.5). For example, the unoriented, "bloblike" RFs evident in V1 and for SAILnet with higher firing rates (Fig. 6.5) are nearly absent from the dictionaries obtained with lower target firing rates. Conversely, for very high firing rates, SAILnet is unable to learn a sparse representation [Zylberberg *et al.*, 2011]. In addition, the overcompleteness of the representation also has a dramatic effect on the receptive field shapes [Rehn and Sommer, 2007; Olshausen *et al.*, 2009; Zylberberg *et al.*, 2011]. For a fixed target firing rate, using a more overcomplete representation results in smaller receptive fields.

The dependence of RF size on mean neuronal firing rate can be understood as follows. The learning rules drive the neural activities to form a linear generative model of the input stimulus, such that the product of each neuron's RF and its firing rate, when summed over all neurons, should match the input. In that case, when the firing rates are low and the neurons are uncorrelated, only a small number of neurons will be active for each stimulus, and thus their receptive fields will need to be fairly large so that the sum of that small number of RFs will span the image space. Conversely, when the firing rates are higher, there may be many neurons active for any given stimulus, and the neurons can subsequently have smaller receptive fields, while still forming a good linear generative model of the stimulus.

6.5 Discussion

I have demonstrated that a computational model can learn V1-like receptive fields while simultaneously exhibiting either a *decrease*, or an *increase*, in the sparseness of neuronal activities. Therefore, the type of active sparseness maximization exhibited by previous sparse coding models [Bell and Sejnowski, 1997; Olshausen and Field,

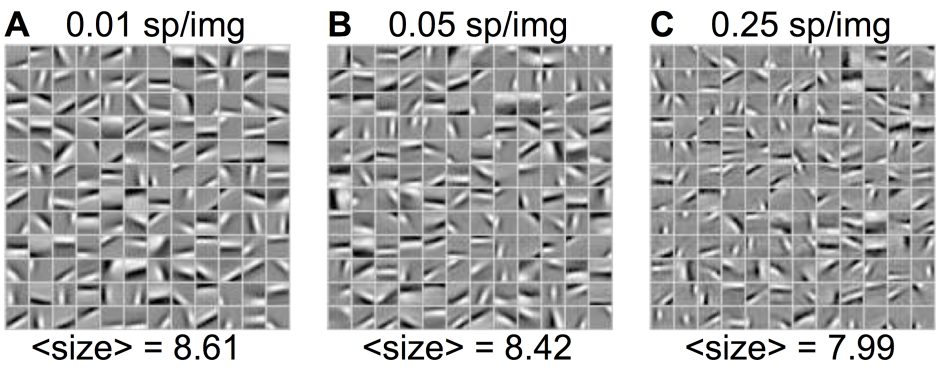

Figure 6.5: **The dictionary of receptive fields learned by SAILnet depends on the homeostatically controlled firing rates.** SAILnet simulations were performed with initial conditions identical to those used to generate the data shown in Fig. 6.4, but with different target firing rates: 0.01, 0.05, or 0.25 spikes per neuron per image. As the target firing rate increases, the RFs become progressively smaller, with more circular "blob-like" shapes emerging. Beneath each panel of RFs, I indicate the average size of the RFs in that dictionary, where I define the size of the j^{th} neuron's RF, s_j, to be $s_j = \sum_i |Q_{ij}| / \sqrt{\sum_i Q_{ij}^2}$. In other words, it is the ratio of the L_1 and L_2 norms of the pixel values in the receptive field. This measure is 1 if there is only one non-zero pixel value in the receptive field, and has a maximum value of $\sqrt{N}$ (where N is the number of pixels in receptive field), when the receptive field uniformly covers the entire image frame. The measure s increases monotonically as one increases the area of support of the receptive field. Thus, it is evident that higher firing rates lead to smaller RFs in the model.

1996; Rehn and Sommer, 2007] is not necessary to produce observed V1 receptive field shapes, nor is it required to learn a sparse representation of natural scenes. Moreover, the SAILnet model is also unique in that it achieves these objectives using only synaptically local information during learning, as required for biological plausibility. Just as important, the computational objective being met by the adaptation process in SAILnet is well defined mathematically [Zylberberg *et al.*, 2011].

In the SAILnet model, homeostatic constraints effectively control the sparseness level. Thus, even though the final state of the network after adaptation must be sparse, if the initial conditions are even sparser than the target level, adaptation decreases the single- and multi-unit sparseness measures over time (Figs. 6.2 and 6.3). Conversely, if the initial conditions are less sparse than the target level, sparseness will increase over time (Fig. 6.4), similar to what is observed in the canonical sparse

coding models such as Sparsenet [Olshausen and Field, 1996] (Fig. 6.1).

I have also found that the specific diversity of RF shapes learned by SAILnet is sensitive to the homeostatic set point for neural firing rates in the model. Similarly, homeostatic mechanisms regulating the total synaptic drive from retina to individual superior colliculus neurons have been implicated in the development of RF shapes in superior colliculus of the mouse [Chandrasekaran *et al.*, 2007], though there are several differences. For example, the stimulus driving activity in that case was composed of spontaneous retinal waves [Butts, 2002], rather than natural visual input, and the main effect reported was a trade-off between the peak firing rate of a neuron in response to punctate stimuli compared with the total area of its RF, rather than qualitative changes in the shapes of RFs as I have found. Homeostatic mechanisms are also reported to play a role in the development of somatosensory maps in rat barrel cortex [Foeller and Feldman, 2004], but the observed changes consist mainly of shifting RF centers rather than qualitative changes in RF shapes.

With the appropriately chosen target firing rate, SAILnet reproduces the range of RF shapes observed in mature V1, including ridge-like edge detectors, Gabor-like filters with multiple subfields, and unoriented "blobs," comparable to the diversity of shapes found by the Sparse Set Coding (SSC) model [Rehn and Sommer, 2007]. Both the SSC model and SAILnet produce a greater diversity of RF shapes as the overcompleteness of the corresponding model is increased — the unoriented shapes in particular are more prevalent in more overcomplete representations.

The developmental trends in sparseness are important because physiology experiments [Imbert and Buisseret, 1975; Cynader *et al.*, 1973] show that receptive fields are a product of developmental processes, yet the majority of theoretical models require that sparseness strictly increases as a result of experience, and thus this type of measurement has the potential to rule out a broad class of models (or at least specific model implementations, including [Olshausen and Field, 1996; Rehn and Sommer, 2007; Bell and Sejnowski, 1997; Smith and Lewicki, 2006]) during development, which is rare in Systems Neuroscience. I hope that this work helps to motivate that experimentation.

The initial motivation for incorporating homeostasis into the SAILnet model was unrelated to the issue of changing sparseness during development. Rather, analytic calculations [Zylberberg *et al.*, 2011] suggested that, in order for synaptically local plasticity rules to optimize the cooperative representation of the stimulus by the neuronal population, the activities would need to remain sparse and decorrelated, and homeostasis is a natural and biologically plausible way to satisfy those conditions. In the retina, for example, spontaneous activity has been found to become less correlated between neuron pairs over time [Demas *et al.*, 2003], broadly consistent with the type of learning rules I am considering. In fact, a variety of homeostatic mechanisms have been reported in various stages of V1 development, including plasticity of intrinsic

excitability as well as synaptic scaling [Turrigiano, 2011]. Homeostatic rules based on synaptic scaling have been incorporated successfully in models of self-organizing maps [Sullivan and de Sa, 2006]. Either of these types of mechanisms could perform the homeostasis modeled by the variable threshold in SAILnet.

These findings suggest that homeostasis, which is widely believed to play a deep role in learning and adaptation by neural systems, may in fact be crucial for the development of the specific diversity of receptive field shapes in visual cortex that can efficiently represent natural images.

The ferret data in Fig. 6.1 were provided by Pietro Berkes.

6.6 Methods

6.6.1 SAILnet simulation details

For all simulations shown herein, the feed-forward weights Q_{ij} were initialized with Gaussian white noise, and the learning rates were set to $(\alpha, \beta, \gamma) = (0.05, 0.002, 0.02)$. For the data shown in Fig. 6.4, the initial recurrent connection strengths W_{ij} were drawn randomly from a $\mathcal{N}(0, 1)$ distribution, the firing thresholds θ_i were initialized to 0.1 $\forall i$, and the model neuronal activity became more sparse during training. For the results shown in Figs. 6.2 and 6.3, the initial recurrent connection strengths were drawn randomly from a $\mathcal{N}(5, 5)$ distribution, the firing thresholds were drawn from a $\mathcal{N}(0.5, 0.25)$ distribution, and the model neuronal activity became less sparse during training.

Sparsenet simulation details

The Sparsenet results shown in Fig. 6.1 were generated using code publicly distributed by Bruno Olshausen (http://redwood.berkeley.edu/bruno/sparsenet/). The "soft" thresholding option was used, and the sparseness parameter λ was set to 0.15, similar to [Olshausen and Field, 1996]. The network was trained with whitened natural images, with feed-forward weights initialized with Gaussian white noise, and the learning rate η was set to 0.25. One cannot achieve decreasing sparseness with Sparsenet by changing the parameters (η, λ), and/or the initial conditions of the feed-forward weights. This can be understood by considering that the rules used to generate Sparsenet model activity explicitly maximize the sparseness. Furthermore, I experimented with several different combinations of parameters and initial conditions, and found in all cases results similar to those shown in Fig. 6.1 (data not shown).

Chapter 7

Conclusions

In this thesis, I have reviewed much literature on natural image statistics, the peripheral mammalian visual system, and the connection therebetween. This connection can be roughly summarized as follows: natural images have statistical regularities that allow for them to be recoding in more information-theoretically efficient ways, and the mammalian visual system reflects many features of this optimal recoding.

While the existing body of work provides an elegant ecological explanation for the observed physiology, explaining both *what* features the visual system should have and *why* those features are beneficial, I have argued that the existing body of research fails to answer the (critical) question of *how* those features should be developed.

Addressing this question forms the main advance from this thesis; below, I discuss the specific contributions made.

7.1 Contributions of this thesis

Sparse coding is an appealing hypothesis that can explain many salient features of the visual cortical physiology. However, the existing sparse coding models fail to be biologically realistic because they employ synaptically non-local plasticity rules, while learning in cortex appears to occur by the modification of the strengths of synaptic connections between neurons, and to depend only on the pre- and post-synaptic activities. In Ch. 5, I demonstrated that, so long as neuronal activities are sparse and uncorrelated, synaptically local plasticity rules suffice to learn a sparse, linear generative code for natural images. I then simulated a model of visual cortex called SAILnet that implemented these ideas, and found that it learned, when trained on natural images, the same diversity of receptive field shapes that are exhibited by simple cells in macaque V1. This provides the first published demonstration that a biophysically plausible theoretical model can actually account for these receptive field

shapes.

Recently, it was observed that, while the theoretical sparse coding models show an increase in sparseness as they adapt to natural image statistics, the sparseness levels in ferret V1 actually decrease as the animal gains visual experience [Berkes *et al.*, 2009] . This is perhaps one of the strongest challenges to the sparse coding hypothesis. However, in Ch. 6, I demonstrated that homeostatic mechanisms, like those in SAILnet, can explain this observation: if the initial conditions are more sparse than the homeostatically regulated set-point, then adaptation leads to decreasing sparseness over time.

Since SAILnet is both a biophysically plausible model for visual cortical development, and the first theoretical model to account for both the observed receptive field shapes, and the developmental trends in sparseness levels, it may be the most parsimonious systems-level explanation for adaptation in the maturing visual cortex.

Finally, with regards to the low-level statistics of natural images, in Ch. 2, I discussed previous work that demonstrated that the observed power-law 2-point functions of natural images could be understood in terms of a collage of randomly placed opaque objects whose sizes follow a power-law distribution. The opacity requirement in the published model, and the observed power-law 2-point functions of medical images, in which objects are more transmissive (less opaque) led us to wonder whether opacity really is an important component of the theoretical understanding. In Ch. 2, I studied a generalization of the previous image model, in which the objects could be made partially transparent, and subsequently demonstrated that, so long as the objects sizes are power-law distributed, opacity does not change the functional form of the 2-point function. This work helps to unify our understanding of the low-level statistics of both natural, and medical images.

7.2 Directions for future research

While the work presented in this thesis represents several important advances, it raises many more issues that form potential directions for future research.

7.2.1 Extend the SAILnet model to the time domain, and model "on-the-fly" inference

One of the most dissatisfying aspects of the SAILnet model, in comparison to biological neural systems, is that SAILnet processes still frames, while real brains view dynamical scenes. Extending SAILnet, and related models, to address this issue would clearly be desirable. However, this problem is not as easily solved as one might think.

In SAILnet, for example, neurons view still frames for an extended period of time, and the neural representation is built up over that period, as the neurons repeatedly spike. If the image keeps changing, and the neural representation is to keep up, then this process of slowly building up the representation must be replaced by a more "on-the-fly", causal inference process in which neurons respond to the past so as to create the present representation, and do so in a relatively short time period, one that is shorter than the time scale over which the stimulus changes significantly. If this speed is not present, then by the time the inference is done, and the stimulus is "understood", the stimulus will have already changed and the animal, while having a decent model of the past visual scene, will not have access to current information about its environment.

This research direction requires confronting the issue of how information might be encoded causally and has been briefly touched upon by [Bialek *et al.*, 1991], but it is still unclear what sort of representation one might expect, and thus what specific functional form to use for the objective function.

One promising approach is suggested by the Liquid State Machine of Mass and colleagues [Maas and Natschlager, 2002], and related ideas of reservoir computing. This approach differs considerably from the efficient coding principles discussed in this thesis, and the inter-relation between these ideas is worthy of study.

7.2.2 What about audition?

Similar to their success in explaining the visual cortex, sparse coding algorithms trained on natural sound ensembles have been shown to successfully reproduce receptive fields in the peripheral auditory system [Smith and Lewicki, 2006]. Like the visual sparse coding problem, Smith & Lewicki's algorithm lacks biophysical plausibility, so a SAILnet-type approach could represent a serious advance in auditory neuroscience. In the specific model of [Smith and Lewicki, 2006], inference was performed acausally: the algorithm was presented with long vectors representing the sound waveforms, and the best set of features to tile the sound clip was chosen, using a "greedy" matching pursuit algorithm. In contrast, an animal processing sounds will not know the future sounds before processing the present ones: at best, it can build up a representation of the immediate past.

Thus, similar to the problem of using SAILnet for dynamical visual stimuli, the solution to this problem requires that we confront the issue of causal, on-the-fly inference, and will thus be a challenge to satisfactorily address.

7.2.3 Understand the connection between neuronal dynamics and inference

In the SAILnet model, we used leaky integrate-and-fire dynamics as a simplified, biologically motivated way, to generate neural outputs from inputs, but made no serious attempt to understand the role that the dynamics play in the inference process, although recent work does suggest that LIF dynamics can find the maximum-likelihood activations in an L_0-minimizing sparse coding model [Shapero *et al.*, 2011].

However, as we have discussed in the introduction to this thesis, real neurons do not strictly follow LIF dynamics, and indeed real neurons are stochastic, such that repeat presentations of the same stimuli lead to different spike trains. This could be the result of slightly different initial conditions on each experimental trial (for example, the electrical potentials and ion concentrations in all of the neurons), or could represent intrinsic neural stochasticity.

With regards to deterministic neuron models, the Hodgkin-Huxley equations provide the most realistic dynamics of any known model, so a natural approach would be to incorporate them into a SAILnet-like model, and to attempt to understand the inference problem being solved by their dynamics.

With regards to intrinsically stochastic neuron models, it has been suggested that stochastic neural dynamics might represent a form of statistical sampling, through which the brain attempts to model the distribution over possible world-states, conditioned on the (noisy) observations it receives. [Buesing *et al.*, 2011]. This work is promising, but its rigorous formulations only strictly apply to simplified neuronal models, and thus understanding how Hodgkin-Huxley dynamics (in the presence of noise) perform inference remains an open (and very interesting) question.

7.2.4 Create models that respect Dale's law

While the model neurons in SAILnet all have direct inhibitory connections between them, excitatory cells in cortex (which the SAILnet neurons are thought to model) can only inhibit each other via the activity of inhibitory interneurons: the excitatory neurons drive interneurons to fire, and the interneurons inhibit the excitatory cells. This is described by Dale's law, which states that every cortical neuron either excites every neuron onto which it synapses, or inhibits them, but that neurons cannot excite some downstream neurons while inhibiting others.

Indeed, in this respect, SAILnet is typical of the existing models of visual cortex, which tend to disregard Dale's law. The creation of a model respecting this constraint is desired because it would allow us to understand how indirect inhibition affects neural computations. To that effect, we are currently developing a model with a separate population of inhibitory neurons. Independent of that work, another group

is pursuing an extension of Rozell's LCA in which there is a separate inhibitory population [Zhu *et al.*, 2012]. These works are complementary, in that Zhu et al. are addressing the inference process, and not the problem of how the network structure might be learned using synaptically local plasticity rules, while our project studies the learning, and not the inference, process.

7.2.5 Move beyond rate coding and linear generative models

As I have implied in the discussion about using SAILnet-like models in the time domain (above), it may be necessary to reconsider how plausible it is to assume that neuronal systems form linear generative models based on firing rates. Indeed, rates cannot be immediately inferred from single-neuron spike trains, instead requiring that one averages over some time window containing several spikes, or averages over several different neurons with near-identical receptive fields [Shadlen and Newsome, 1994].

At the same time, recent work has cast in doubt the plausibility of rate coding by showing that a Bayesian-optimal decision algorithm performs worse than a real animal in a visual 2-alternative forced-choice task when that algorithm uses retinal ganglion cell firing rates as its inputs, but that the Bayes-optimal algorithm can match the animal's performance when it uses spike times instead [Jacobs *et al.*, 2009]. Since the Bayes-optimal decoder puts an upper limit on the performance possible with rate coding, this work suggests that we explore other codes, like timing codes, to model neural systems.

Towards this end, some recent work has explored the notion of sparse coding without requiring linear generative models, in which the neural representation is made to maximize the mutual information between itself and the stimulus without ever specifying how that representation would be decoded [Karklin and Simoncelli, 2011]. That work represents an important step in the right direction, but it still uses continuous-valued outputs from each of the individual computational units (analogous to something like firing rates), and made no attempt at biophysical plausibility of either the learning, or the inference, steps in the algorithm.

There is thus an opportunity for research that solves these problems, to understand how neuronal representations might be formed without appealing to rates and linear decoding.

7.2.6 Summary

In the preceding few subsections, I have suggested several different lines of inquiry that would all provide meaningful extensions to the work presented in the earlier chapters of this thesis. These generally involve the addition of more biophysical realism to the theoretical models. However, I wish to point out here that highly realistic simulations

on their own are not very useful unless their functions are deeply understood. In the extreme case, imagine creating a perfectly realistic simulation of a brain, in which we simply put every known anatomical and physiological fact into a giant computer, and run the simulation, similar to the blue brain project (http://bluebrain.epfl.ch/). In that case, the simulation has so much complexity that understanding which of its details contribute to each specific function would not necessarily be much easier than attempting to get the same understanding from experimenting directly on real brains, although lesion experiments, in which features are iteratively removed from the simulation to understand the deficits caused by their removal, could potentially be performed more quickly in the simulated brain than in real brains.

In my opinion, a better approach is to gradually add in more biophysical realism to the models, and at each stage to understand how the new features work, and how they affect the function of the whole system. In so doing, once the "final" model is built, that contains all of the important pieces, we will actually understand that model well enough to use it to make concrete statements about biology. The technical work in this thesis represents one step in this chain, in which I elucidated the conditions under which synaptically local learning rules can optimize the cooperative neural representation, and studied a model that implements this concept. The future research directions I proposed in this last chapter pick up that torch, and carry on in the same spirit.

Bibliography

[Abbott and Nelson, 2000] L.F. Abbott and S.B. Nelson. Synaptic plasticity: taming the beast. *Nat. Neurosci.*, 3:1178–1183, 2000.

[Abeles *et al.*, 1990] M. Abeles, E Vaadia, and H. Berman. Firing patterns of single units in the prefrontal cortex and neural network models. *Network: comput. neural systems*, 1:13– 25, 1990.

[Adams and Horton, 2002] D.L. Adams and J.C. Horton. Shadows cast by retinal blood vessels mapped in primary visual cortex. *Science*, 298:572–576, 2002.

[Atick and Redlich, 1992] J. J. Atick and A. N. Redlich. What does the retina know about natural scenes. *Neural Comput.*, 4:196–210, 1992.

[Attneave, 1954] F. Attneave. Some informational aspects of visual psychology. *Psychol. Rev.*, 61:183–193, 1954.

[Baddeley *et al.*, 1997] R. Baddeley, L. F. Abbott, M. C. A. Booth, F. Sengpiel, T. Freeman, et al. Responses of neurons in primary and inferior temporal visual cortices. *Proc. R Soc. Lon. B*, 264:1775– 1783, 1997.

[Balboa and Grzywacz, 2003] R.M. Balboa and N.M. Grzywacz. Power spectra and distribution of contrasts of natural images from different habitats. *Vision Res.*, 43:2527–2537, 2003.

[Balboa *et al.*, 2001] R.M. Balboa, C.W. Tyler, and N.M Grzywacz. Occlusions contribute to scaling in natural images. *Vision Res.*, 41:955–964, 2001.

[Barlow, 1961] H. B. Barlow. Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. Cambridge: MIT Press, 1961.

[Bell and Sejnowski, 1997] A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Res.*, 37:3327–3328, 1997.

[Berkes *et al.*, 2009] P. Berkes, B. L. White, and J. Fiser. No evidence for active sparsification in the visual cortex. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 108–116. San Mateo: Morgan Kaufmann, 2009.

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

[Bialek *et al.*, 1991] W. Bialek, F. Rieke, R. R. de Ruyter Van Steveninck, and D. Warland. Reading a neural code. *Science*, 252:1854– 1857, 1991.

[Bialek *et al.*, 1993] W. Bialek, M. DeWeese, F. Rieke, and D. Warland. Bits and brains: information flow in the nervous system. *Physica A*, 200:581–592, 1993.

[Bordenave *et al.*, 2006] C. Bordenave, Y. Gousseau, and F. Roueff. The dead leaves model: a general tessellation modeling occlusion. *Adv. Appl. Prob.*, 38:31–46, 2006.

[Buesing *et al.*, 2011] L. Buesing, J. Bill, B. Nessler, and W. Mass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comp. Biol.*, 7:e1002211–1 – e1002211–22, 2011.

[Burrone and Murthy, 2003] J. Burrone and V. N. Murthy. Synaptic gain control and homeostasis. *Curr. Opin. Neurobiol.*, 13:560–567, 2003.

[Butts, 2002] D.A. Butts. Retinal waves: implications for synaptic learning rules during development. *Neuroscientist*, 8:243–253, 2002.

[Carandini *et al.*, 1997] M. Carandini, D.J. Heeger, and J.A. Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.*, 17:8621–8644, 1997.

[Carlson, 1978] C.R. Carlson. Thresholds for perceived image sharpness. *Photographic Sci. and Eng.*, 22:69–71, 1978.

[Chandrasekaran *et al.*, 2007] A.R. Chandrasekaran, R.D. Shah, and M.C. Crair. Developmental homeostasis of mouse retinocollicular synapses. *J. Neurosci.*, 27:1746–1755, 2007.

[Clopath *et al.*, 2010] C. Clopath, L. Büsing, E Vasilaki, and W Gerstner. Connectivity reflects coding: a model of voltage-based stdp with homeostasis. *Nat. Neurosci.*, 13:344– 352, 2010.

[Cover and Thomas, 1991] T.M. Cover and J.A. Thomas. *Elements of information theory*. New York: Wiley & Sons, 1991.

[Curcio and Allen, 1990] C.A. Curcio and K.A. Allen. Topography of ganglion cells in human retina. *J. Comp. Neurol.*, 300:5–25, 1990.

[Cynader *et al.*, 1973] M. Cynader, N. Berman, and A. Hein. Cats reared in stroboscopic illumination: effects on receptive fields in visual cortex. *Proc. Natl. Acad. Sci. USA*, 70:1353–1354, 1973.

[Dan and Poo, 2006] Y. Dan and M. M. Poo. Spike timing-dependent plasticity: from synapse to perception. *Physiol. Rev.*, 86:1033– 1048, 2006.

[Dan *et al.*, 1996] Y. Dan, J. J. Atick, and C. R. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.*, 16:3351–3362, 1996.

[Davis and Bezprozvanny, 2001] G.W. Davis and I. Bezprozvanny. Maintaining the stability of neural function: a homeostatic hypothesis. *Ann. Rev. Physiol.*, 63:847–869, 2001.

[Dayan and Abbott, 2001] P. Dayan and L. Abbott. *Theoretical neuroscience: computational and mathematical modelling of neural systems*. Cambridge: MIT Press, 2001.

[Delorme *et al.*, 2000] A. Delorme, L. Perrinet, and S. J. Thorpe. Networks of integrate-and-fire neurons using rank order coding b: spike timing dependent plasticity and emergence of orientation selectivity. *Neurocomput.*, 38:539– 545, 2000.

[Demas *et al.*, 2003] J. Demas, S.J. Eglen, and R.O. Wong. Developmental loss of synchronous spontaneous activity in the mouse retina is independent of visual experience. *J. Neurosci.*, 23:2851–2860, 2003.

[Desai *et al.*, 1999] N.S. Desai, L.C. Rutherford, and G.G Turrigiano. Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.*, 2:515–520, 1999.

[DeWeese, 1996] M. DeWeese. Optimization principles for the neural code. *Network: comput. neural systems*, 7:325–331, 1996.

[Dong and Atick, 1995a] D.W. Dong and J.J. Atick. Statistics of natural time-varying images. *Network: Comput. Neural Systems*, 6:345–358, 1995.

[Dong and Atick, 1995b] D.W. Dong and J.J. Atick. Temporal decorrelation: a theory of lagged and non lagged responses in the lateral geniculate nucleus. *Network: Comput. Neural Systems*, 6:159–178, 1995.

[Donoho, 2004] D. L. Donoho. Compressed sensing. *IEEE trans. inform. theory*, 52:1289–1396, 2004.

[Ecker *et al.*, 2010] A.S. Ecker, P. Berens, G. A. Keliris, M. Bethge, N. K. Logothetis, et al. Decorrelated neuronal firing in cortical microcircuits. *Science*, 327:584– 587, 2010.

[Engel *et al.*, 1997] S.A. Engel, G.H. Glover, and B.A. Wandell. Retinotopic organization in human visual cortex and the spatial precision of functional mri. *Cereb. Cortex*, 7:181–192, 1997.

[Falconbridge *et al.*, 2006] M. S. Falconbridge, R. L. Stamps, and D. R. Badcock. A simple hebbian/anti-hebbian network learns the sparse, independent components of natural images. *Neural Comput.*, 18:415– 429, 2006.

[Feldman, 2009] D.E. Feldman. Synaptic mechanisms for plasticity in neocortex. *Annu. Rev. Neurosci.*, 32:33–55, 2009.

[Field, 1987] D.J. Field. Relation between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4:2379–2394, 1987.

[Foeller and Feldman, 2004] E. Foeller and D.E. Feldman. Synaptic basis for developmental plasticity in somatosensory cortex. *Curr. Opin. Neurobiol.*, 14:89–95, 2004.

[Földiák, 1990] P. Földiák. Forming sparse representations by a local anti-hebbian rule. *Biol. Cybern.*, 64:165–170, 1990.

[Gaese and Ostwald, 2003] B. H. Gaese and J. Ostwald. Complexity and temporal dynamics of frequency coding in the awake rat auditory cortex. *Eur. J. Neurosci.*, 18:2638– 2652, 2003.

[Garrigues and Olshausen, 2008] P. J. Garrigues and B. A. Olshausen. Learning horizontal connections in a sparse coding model of natural images. In J. C. Platt, D. Kollar, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 505–512. San Mateo: Morgan Kaufmann, 2008.

[Gawne *et al.*, 1996] T.J. Gawne, T.W. Kjaer, J.A. Hertz, and B.J. Richmond. Adjacent visual cortical complex cells share about 20% of their stimulus-related information. *Cereb. Cortex*, 6:482–489, 1996.

[Gutnisky and Dragoi, 2008] D.A. Gutnisky and V. Dragoi. Adaptive coding of visual information in neural populations. *Nature*, 452:220–224, 2008.

[Haider *et al.*, 2010] B. A. Haider, M. R. Krause, A. Duque, Y. Yu, J. Touryan, et al. Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron*, 65:107–121, 2010.

[Heine and Velthuizen, 2002] J.J. Heine and R.P. Velthuizen. Spectral analysis of full field digital mammography data. *Med. Phys.*, 29:647–661, 2002.

[Hendry and Jones, 1988] S.H.C. Hendry and E.G. Jones. Activity-dependent regulation of gaba expression in the visual cortex of adult monkeys. *Neuron*, 1:701–712, 1988.

[Hodgkin and Huxley, 1952] A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117:500–544, 1952.

[Hopfield, 1984] J. Hopfield. Neurons with graded responses have collective properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA*, 81:3088– 3092, 1984.

[Hromádka *et al.*, 2008] T. Hromádka, M. R. DeWeese, and A. M. Zador. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol.*, 6:124–137, 2008.

[Hsiao and Millane, 2005] W.H. Hsiao and R.P. Millane. Effects of occlusion, edges, and scaling on the power spectra of natural images. *J. Opt. Soc. Am. A*, 22:1789–1797, 2005.

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

[Imbert and Buisseret, 1975] M. Imbert and P. Buisseret. Receptive field characteristics and plastic properties of visual cortical cells in kittens reared with or without visual experience. *Exp. Brain Res.*, 22:25–36, 1975.

[Izhikevich, 2007] E.M. Izhikevich. *Dynamical systems in neuroscience: the geometry of excitability and bursting*. Cambridge: MIT Press, 2007.

[Jacobs *et al.*, 2009] A.L. Jacobs, G. Fridman, R.M. Douglas, N.M. Alam, P. Latham, G.T. Prusky, and S. Nirenberg. Ruling out and ruling in neural codes. *Proc. Natl. Acad. Sci. USA*, 14:5936–5941, 2009.

[Karklin and Simoncelli, 2011] Y. Karklin and E.P. Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS*11)*, volume 24. MIT Press, 2011.

[Klekamp *et al.*, 1991] J. Klekamp, A. Riedel, C. Harper, and H.J. Kretschmann. Quantitative changes during the postnatal maturation of the human visual cortex. *J. Neurol. Sci.*, 103:136–143, 1991.

[Kohn and Smith, 2005] A. Kohn and M.A. Smith. Stimulus dependence of neuronal correlation in primary visual cortex. *J. Neurosci.*, 25:3661–3673, 2005.

[Kolb *et al.*, 2011] H. Kolb, R. Nelson, E. Fernandez, and B. Jones. Webvision: the organization of the retina and visual system. [`http://webvision.med.utah.edu/`; accessed April 2012], 2011.

[Koulakov *et al.*, 2009] A. A. Koulakov, T. Hromádka, and A. M. Zador. Correlated connectivity and the distribution of firing rates in the neocortex. *J. Neurosci.*, 29:3685– 3694, 2009.

[Lee *et al.*, 2001] A.B. Lee, D. Mumford, and J. Huang. Occlusion models for natural images: a statistical study of a scale-invariant dead leaves model. *Int. J. Comp. Vis.*, 41:35–59, 2001.

[Leuba and Kraftsik, 1994] G. Leuba and R. Kraftsik. Changes in volume, surface estimate, three-dimensional shape and total number of neurons in the human primary visual cortex from midgestation until old age. *Anat. Embryol.*, 190:351–366, 1994.

[Linsker, 1986] R. Linsker. An application of the principle of maximum information preservation to linear systems. In D.S. Touretzky, editor, *Advances in Neural Information Processing 1*, pages 186–194. San Mateo: Morgan Kaufmann, 1986.

[London *et al.*, 2010] M. London, A. Roth, L. Beeren, M. Häuser, and P.E. Latham. Sensitivity to perturbations *in vivo* implies high noise and suggests rate coding in cortex. *Nature*, 446:123–127, 2010.

[Maas and Natschlager, 2002] W. Maas and T. Natschlager. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.*, 14:2351–2560, 2002.

[Mackay, 2003] D.J.C. Mackay. *Information theory, inference, and learning algorithms*. New York: Cambridge university press, 2003.

[Mainen and Sejnowski, 1995] Z. F. Mainen and T. J. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 286:1503– 1506, 1995.

[Marder and Goaillard, 2006] E. Marder and J.M. Goaillard. Variability, compensation and homeostasis in neuron and network function. *Nat. Rev. Neurosci.*, 7:563–574, 2006.

[Masquelier and Thorpe, 2007] T. Masquelier and S. J. Thorpe. Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol.*, 3:e31, 2007.

[Matheron, 1968] G. Matheron. Modèle séquentiel de partition aléatoire, 1968. Technical report of le Centre de Morphologie Mathématique, Fontainebleau.

[Mrsic-Flogel *et al.*, 2007] T.D. Mrsic-Flogel, S.B. Hofer, K. Ohki, R.C. Reid, T. Bonhoeffer, and M. Hübener. Homeostatic regulation of eye-specific responses in visual cortex during ocular dominance plasticity. *Neuron*, 54:961–972, 2007.

[Oja, 1982] E. Oja. A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, 15:267– 273, 1982.

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

[Olshausen and Field, 1997] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Res.*, 37:3311–3325, 1997.

[Olshausen *et al.*, 2009] B. A. Olshausen, C. F Cadieu, and D. K. Warland. Learning real and complex overcomplete representations from the statistics of natural images. *Proc. SPIE*, 7446:74460S–1 – 74460S–11, 2009.

[Perrinet *et al.*, 2004] L. Perrinet, M. Samuelides, and S. Thorpe. Sparse spike coding in an asynchronous feed-forward multi-layer neural network using matching pursuit. *Neurocomput.*, 57:125– 34, 2004.

[Perrinet, 2010] L. Perrinet. Role of homeostasis in learning sparse representations. *Neural Comput.*, 22:1812– 1836, 2010.

[Pillow *et al.*, 2008] J.W. Pillow, J. Shlens, L. Paninski, A. Sher, A.M. Litke, E.J. Chichilnisky, and E.P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454:995–999, 2008.

[Pitkow and Meister, 2012] X. Pitkow and M. Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nat. Neurosci.*, 15:628–635, 2012.

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

[Reich *et al.*, 2001] D.S. Reich, F. Mechler, and J.D. Victor. Independent and redundant information in nearby cortical neurons. *Science*, 294:2566–2568, 2001.

[Renart *et al.*, 2010] A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, et al. The asynchronous state in cortical circuits. *Science*, 327:587– 590, 2010.

[Ringach, 2002] D. Ringach. Spatial structure and asymmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.*, 88:455–463, 2002.

[Rochefort *et al.*, 2009] N. L. Rochefort, O. Garaschuk, R. Milos, M. Narushima, N. Marandi, B. Pichler, Y. Kovalchuk, and A. Konnerth. Sparsification of neuronal activity in the visual cortex at eye-opening. *Proc. Natl. Acad. Sci. USA*, 106:15049– 15054, 2009.

[Rozell *et al.*, 2008] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.*, 20:2526– 2563, 2008.

[Ruderman and Bialek, 1994] D. Ruderman and W. Bialek. Statistics of natural images: scaling in the woods. *Phys. Rev. Lett.*, 73:814–817, 1994.

[Ruderman, 1997] D.L. Ruderman. Origins of scaling in natural images. *Vision Res.*, 37:3385–3398, 1997.

[Savin *et al.*, 2010] S. Savin, P. Joshi, and J. Triesch. Independent component analysis in spiking neurons. *PLoS Comput. Biol*, 6:e1000757, 2010.

[Schneidman *et al.*, 2006] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006.

[Shadlen and Newsome, 1994] M.N. Shadlen and W.T. Newsome. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.*, 4:569–579, 1994.

[Shannon, 1948] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[Shapero *et al.*, 2011] S. Shapero, D. Brüderle, P. Hasler, and C. Rozell. Sparse approximation on a network of locally competitive integrate and fire neurons, 2011. CoSyNe 2011 abstract.

[Sharpee *et al.*, 2004] T. Sharpee, N.C. Rust, and W. Bialek. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput.*, 16:223–250, 2004.

[Sharpee *et al.*, 2006] T. Sharpee, H. Sugihara, A.V. Kurgansky, S. P. Rebrik, M.P. Stryker, and K.D. Miller. Adaptive filtering enhances information transmission in visual cortex. *Nature*, 439:936–942, 2006.

[Shepherd, 2004] G.M. Shepherd. *The synaptic organization of the brain,* 5^{th} *ed.* New York: Oxford university press, 2004.

[Simoncelli and Olshausen, 2001] E.P. Simoncelli and B.A. Olshausen. Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, 24:1193–1216, 2001.

[Smith and Kohn, 2008] M.A. Smith and A. Kohn. Spatial and temporal scales of neuronal correlation in primary visual cortex. *J. Neurosci.*, 28:12591–12603, 2008.

[Smith and Lewicki, 2006] E.C. Smith and M.S. Lewicki. Efficient auditory coding. *Nature*, 439:978–982, 2006.

[Song. *et al.*, 2005] S. Song., P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.*, 3:0507– 0519, 2005.

[Stanley *et al.*, 1999] G. Stanley, F.F. Li, and Y. Dan. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J. Neurosci*, 19:8036–8042, 1999.

[Sullivan and de Sa, 2006] T.J. Sullivan and V.R. de Sa. Homeostatic synaptic scaling in self-organizing maps. *Neural Netw.*, 19:734–743, 2006.

[Tolhurst *et al.*, 2009] D. J. Tolhurst, D. Smyth, and I. D. Thompson. The sparseness of neuronal responses in ferret primary visual cortex. *J. Neurosci.*, 29:2355–2370, 2009.

[Torralba and Oliva, 2003] A. Torralba and A. Oliva. Statistics of natural images categories. *Network: Comput. Neural Systems*, 14:391–412, 2003.

[Treves and Rolls, 1991] A. Treves and E. T. Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Comput. Neural Syst*, 2:371–397, 1991.

[Turrigiano and Nelson, 2000] G.G. Turrigiano and S.B. Nelson. Hebb and homeostasis in neuronal plasticity. *Curr. Opin. Neurobiol.*, 10:358–364, 2000.

[Turrigiano *et al.*, 1998] G.G. Turrigiano, K.R. Leslie, N.S. Desai, L.C. Rutherford, and S.B. Nelson. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391:892–896, 1998.

[Turrigiano, 2011] G. Turrigiano. Too many cooks? intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annu. Rev. Neurosci.*, 34:89–103, 2011.

[VanRullen and Thorpe, 2002] R. VanRullen and S. J. Thorpe. Surfing a spike wave down the ventral steam. *Vision Res.*, 42:2593– 2615, 2002.

[Venter *et al.*, 2001] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Murray, G.G. Sutton, H.O. Smith, et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

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

[Vinje and Gallant, 2002] W. E. Vinje and J. L. Gallant. Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1. *J. Neurosci.*, 22:2904–2915, 2002.

[Wassle *et al.*, 1981] H. Wassle, B.B. Boycott, and R.-B. Illing. Morphology and mosaic of on- and off- beta cells in the cat retina and some functional considerations. *Proc. R. Soc. Lond. B*, 212:177–195, 1981.

[White and Fitzpatrick, 2007] L.E. White and D. Fitzpatrick. Vision and cortical map development. *Neuron*, 56:327–338, 2007.

[Zhu *et al.*, 2012] M. Zhu, B. Olshausen, and C. Rozell. Biophysically accurate inhibitory interneuron properties in a sparse coding network, 2012. CoSyNe 2012 abstract.

[Zylberberg *et al.*, 2011] J. Zylberberg, J. T. Murphy, and M. R. DeWeese. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS Comp. Biol.*, 7:e1002250–1 – e1002250–12, 2011.