SPARSE CODING MODELS OF NEURAL RESPONSE IN THE PRIMARY VISUAL CORTEX

A Dissertation
Presented to
The Academic Faculty

By

Mengchen Zhu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Biomedical Engineering

Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology
August 2015

SPARSE CODING MODELS OF NEURAL RESPONSE IN THE PRIMARY VISUAL CORTEX

Approved by:

Dr. Christopher J. Rozell, Advisor
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Garrett B. Stanley
Professor, Department of BME
Georgia Institute of Technology

Dr. Bruno A. Olshausen
Professor, Helen Wills Neuroscience Institute
UC Berkeley

Dr. Robert J. Butera
Professor, Department of BME
Georgia Institute of Technology

Dr. Ilya Nemenman
Associate Professor, Department of Physics
Emory University

Date Approved: April 16, 2015

ACKNOWLEDGMENT

I am deeply indebted to my advisor Christopher Rozell for his unfailing support during the past few years. This work would be impossible without his guidance and feedback. I would like to thank Bruno Olshausen for helpful discussions that shaped many aspects of this work. I would like to acknowledge Ian Stevenson, Urs Köster, Adam Charles, and Sam Shapero for a productive collaboration. I am also grateful to Amir Khosrowshahi and Charlie Gray for their help during the initial phase of this work. I am extremely fortunate to have worked with many wonderful people in the Neurolab and CSIP at Georgia Tech, including Allie Del Giorno, Han-Lun Yap, Doug Ollerenshaw, Daniel Millard, Clarissa Shephard, Sean Kelly, Darryl Sale, Steve Conover, Abbie Kressner, Aurele Balavoine, Nick Bertrand, Marissa Norko, Pavel Dunn, Michael Moore, Andrew Massimino, and many others. This work was made possible through a NIH CRCNS grant R01EY019965.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Sparse coding is an influential unsupervised learning approach proposed as a theoretical model of the encoding process in the primary visual cortex (V1). While sparse coding has been successful in explaining classical receptive field properties of simple cells, it was unclear whether it can account for more complex response properties in a variety of cell types. In this dissertation, we demonstrate that sparse coding and its variants are consistent with key aspects of neural response in V1, including many contextual and nonlinear effects, a number of inhibitory interneuron properties, as well as the variance and correlation distributions in the population response. The results suggest that important response properties in V1 can be interpreted as emergent effects of a neural population efficiently representing the statistical structures of natural scenes under resource constraints. Based on the models, we make predictions of the circuit structure and response properties in V1 that can be verified by future experiments.

CHAPTER I

INTRODUCTION

We are still far from understanding visual processing in the primary visual cortex (V1) in behaviorally relevant settings [17]. While the classical feedforward receptive field framework pioneered by Hubel and Wiesel [18] describes how V1 neurons selectively respond to artificial visual stimuli, it does not answer in what way this response is relevant to perception in a naturalistic environment. In contrast to this descriptive approach, sparse coding theory offers a normative alternative and asks the following question: assuming that the visual system is evolved/adapted to represent natural stimuli efficiently and faithfully (so that the stimuli can be reconstructed/decoded accurately), what neural representation do we expect? Remarkably, with unsupervised learning from natural images, the model self-adapts to the low-dimensional structures in images and forms filters that resemble the classical receptive fields in simple cells described by Hubel and Wiesel [12, 14], suggesting that sparse coding may indeed be a strategy employed by V1. While highly influential in the past two decades, sparse coding as a model for V1 has been largely restricted to one type of response property for one cell type. In this dissertation we examine whether sparse coding can go beyond the classical receptive field properties in simple cells by testing if its prediction is consistent with other aspects of visual encoding. In particular we try to answer three questions:

1. Can sparse coding account for important response nonlinearities and contextual effects in V1 simple cells?

2. Can biologically plausible inhibitory interneuron population be incorporated into a neural network implementation of sparse coding?

3. Can sparse coding account for population statistics of neural response to dynamic

natural scenes?

Answers to these questions have important biological implications. Indeed, these questions all concern essential yet not well-understood aspects of visual encoding in V1. First of all, despite being well-documented since Hubel and Wiesel's original study [19], contextual effect's role in visual coding remains unclear. Secondly, while inhibitory interneurons are increasingly identified as a dominant force in shaping cortical activities [20], how they interact with the excitatory population to encode the stimulus is an ongoing active area of research. Finally, with the advent of large scale multi-electrode recordings, there is a mismatch between the amount of population recording data generated and how little we understand the population activities. Importantly, we currently lack insights into how distributions of the population response reflect the underlying encoding process [21].

From a modeling perspective, the present work represents a step towards developing biologically verifiable coding models. Efficient coding hypothesis exemplified by sparse coding has been proposed as a principle for understanding sensory system and its relation with the environment. While greatly influential, this hypothesis suffers a lack of clear biological definition of model output and a consequent difficulty with experimental validation [22]. In this work, we construct biologically plausible network implementations of sparse coding based on a dynamical system model [23]. This not only makes model outputs directly comparable with biological observations, but also generates predictions on biological network structures and properties that can be tested in future experiments.

While motivated by reproducing biology, models developed in this dissertation in several cases demonstrate computational advantages over the original sparse coding. A case in point: in Chap. 4, introducing an alternative constraint that encourages more biologically realistic response distribution in sparse coding also produces a "better" code in terms of sparsity and reconstruction error. These computationally more favorable models may find further applications in signal processing and computer vision where sparse coding has proven useful.

Taken together, characterizing and modeling different aspects of visual processing in V1 under a unifying computational framework represents a step towards a principled understanding of population encoding. The success of this approach implies that many disparate response properties in V1 are emergent characteristics of a neural population adapted to represent natural stimuli efficiently.

The dissertation is organized as follows. In Chap. 2, we investigate the first question regarding contextual and nonlinear effects. Simple cells in V1 demonstrate many response properties that are either nonlinear or involve response modulations (i.e., stimuli that do not cause a response in isolation alter the cell's response to other stimuli). These non-classical receptive field (nCRF) effects are generally modeled individually and their collective role in biological vision is not well understood. In this chapter, we perform extensive simulated physiology experiments to show that many nCRF response properties are simply emergent effects of a dynamical system implementing the sparse coding model. These results suggest that rather than representing disparate information processing operations themselves, these nCRF effects could be consequences of an optimal sensory coding strategy that attempts to represent each stimulus most efficiently. This interpretation provides a potentially unifying high-level functional interpretation to many response properties that have generally been viewed through distinct models.

In Chap. 3, we study the second question about inhibitory interneurons. Cortical function is a result of coordinated interactions between excitatory and inhibitory neural populations. In previous theoretical models of sensory systems, inhibitory neurons are often ignored or modeled too simplistically to contribute to understanding their role in cortical computation. In biophysical reality, inhibition is implemented with interneurons that have different characteristics from the population of excitatory cells. In this chapter, we propose a computational approach for including inhibition in theoretical models of neural coding in a way that respects several of these important characteristics, such as the relative number of inhibitory cells and the diversity of their response properties. The main idea is that the

significant structure of the sensory world is reflected in very structured models of sensory coding, which can then be exploited in the implementation of the model using modern computational techniques. We demonstrate this approach on sparse coding that has been successful at modeling other aspects of sensory cortex.

In Chap. 4, we examine the third question concerning the population statistics of response to natural scenes. Natural vision is a result of coordinated activity of populations of neurons. To understand this population activity, we characterize the statistical distribution of the population response in multi-electrode recordings of animals viewing natural movies. We then compare the sparse coding model predicted response distribution with the measured distribution. Interestingly, while capturing the overall spike rate distribution, sparse coding does not predict the prevalence of silent neurons and clustered correlation in the data. To address this model inadequacy, we incorporate additional constraints into the model, which gives rise to model distribution much closer to the biological distribution. In addition, these added constraints result in a computationally more efficient and invariant visual code.

CHAPTER II

VISUAL NONCLASSICAL RECEPTIVE FIELD EFFECTS EMERGE FROM SPARSE CODING IN A DYNAMICAL SYSTEM[1]

Extensive electrophysiology studies have shown that many V1 simple cells have nonlinear response properties to stimuli within their classical receptive field (CRF) and receive contextual influence from stimuli outside the CRF modulating the cell's response. Models seeking to explain these non-classical receptive field (nCRF) effects in terms of circuit mechanisms, input-output descriptions, or individual visual tasks provide limited insight into the functional significance of these response properties because they do not connect the full range of nCRF effects to optimal sensory coding strategies. The (population) sparse coding hypothesis conjectures an optimal sensory coding approach where a neural population uses as few active units as possible to represent a stimulus. We demonstrate that a wide variety of nCRF effects are emergent properties of a single sparse coding model implemented in a neurally plausible network structure (requiring no parameter tuning to produce different effects). Specifically, we replicate a wide variety of nCRF electrophysiology experiments (e.g., end-stopping, surround suppression, contrast invariance of orientation tuning, cross-orientation suppression, etc.) on a dynamical system implementing sparse coding, showing that this model produces individual units that reproduce the canonical nCRF effects. Furthermore, when the population diversity of an nCRF effect has also been reported in the literature, we show that this model produces many of the same population characteristics. These results show that the sparse coding hypothesis, when coupled with a biophysically plausible implementation, can provide a unified high-level functional interpretation to many response properties that have generally been viewed through distinct mechanistic or phenomenological models.

[1] Results presented here were previously published in [24]

2.1 Introduction

As we seek to understand how sensory nervous systems process information about their environment, one of the most common quantitative descriptors of neural coding has been the notion of a classical receptive field (CRF) [25]. In general, the CRF is a measurement of the portion of the stimulus space that causes a change in a neuron's response when a stimulus is presented (or removed). For example, beginning with the pioneering work of Hubel and Wiesel [26], simple cells in the primary visual cortex (V1) have been characterized as feature detectors with CRFs that are selective for location, orientation and spatial frequency.

Unfortunately, a simple linear-nonlinear model based on the measured CRF (e.g., linear filtering with the CRF followed by nonlinear thresholding or saturation) is insufficient to explain many response properties of V1 cells. For example, extensive electrophysiology studies have shown that many V1 simple cells also receive contextual influence where stimuli not part of the CRF can modulate the cell's response to CRF stimuli (reviewed in [27]). Furthermore, when driven by rich stimuli within the CRF, simple cells exhibit complex nonlinear response properties that cannot be captured by thresholding or saturation alone [28]. We use the term *non-classical receptive field (nCRF) effects* to collectively refer to these contextual modulations and nonlinear response properties.

Understanding nCRF effects is likely critical for understanding the coding of natural stimuli because they arise under stimulus conditions that are more complex and ecologically relevant than the stimuli often used in CRF mapping experiments (e.g., sinusoidal gratings, white noise, sparse dots). Indeed, recent electrophysiology experiments with natural video stimuli have shown contextual influence in V1 responses [29–32]. Furthermore, observed V1 nCRF effects have been related to perceptual contextual effects such as contour integration [33].

Given the wide range of different nCRF effects reported in the literature, it is still unclear how these effects are related or what collective role they play in sensory coding.

Many individual nCRF effects have been successfully described in terms of potential underlying circuit mechanisms (i.e., mechanistic models, reviewed in [34]) or compact stimulus/response descriptions (i.e., phenomenological models, reviewed in [27]). While valuable, these approaches do not fully address the functional significance of nCRF effects or illuminate their role in sensory information processing. In another direction, individual nCRF effects have also been connected to potential benefits in specific tasks (e.g., curvature detection [35], contour integration as reviewed in [36], figure-ground segregation as reviewed in [37]). While these approaches are also valuable, these types of models have limited explanatory power because they only address narrow subsets of biological vision (i.e., individual tasks) and they do not show that the processing strategies represented by nCRF effects are optimal for the given tasks. In short, models constructed for individual effects do not connect this broad range of response properties to the optimal sensory coding strategies that can provide a parsimonious description in terms of fundamental system goals.

One central goal of theoretical and computational biology is to provide functional insight into biological phenomenon by using high-level models (often abstracting away specific experimental detail) to generalize and explain disparate observations. Regarding CRF properties in biological vision, one model that has had success in this regard is the sparse coding hypothesis. Sparse coding conjectures an optimal coding goal where a population of cells encodes a stimulus at a given time using as few active units as possible. Specifically, the model of interest optimizes *population* sparsity, which is distinct from *lifetime* sparsity (a single cell being active a small fraction of the time). In seminal results, the high-level sparse coding model (combined with unsupervised learning using the statistics of natural images) has been shown to be sufficient to explain the emergence of V1 CRF shapes both qualitatively [12] and quantitatively [38]. In addition to this success providing functional insight into CRF properties, distributed sparse neural codes have many potential

benefits (e.g., explicit information representation and easy decodability at higher processing stages [39], metabolic efficiency [40], increased capacity of associative and sequence memory models [41, 42]) and are consistent with many recent electrophysiology experiments [43].

Despite the success accounting for the emergence of CRF properties, there has been little work showing that sparse coding can account for response properties observed in V1 cells. There have been several recent experimental results showing that stimuli in the CRF surround can cause individual cell responses with higher lifetime sparsity than expected (e.g. [29,30,32], reviewed in [44]). While this experimental observation provides encouraging support for the sparse coding hypothesis, it does not imply that a sensory coding model optimizing sparsity is sufficient to account for V1 response properties (including nCRF effects). Sparse coding is one interpretation of the efficient coding hypothesis [22] (conjecturing that neural coding should successively remove stimulus redundancy), and other models related to efficient coding have shown individual model cells that produce some nCRF effects (reviewed in detail in the Discussion section). However, few of these models have shown the broad spectrum of observed nCRF effects in single cells, and none have yet demonstrated the diversity of population response properties reported in the literature for a single effect. Taken together, the evidence of sparsity in experimental observations and the prior success of other related models gives motivation for investigating the potential role of sparse coding in producing nonclassical response properties.

In this chapter we demonstrate that a wide variety of nCRF effects are emergent properties of a sparse coding model implemented in a neurally plausible network structure. Specifically, we use the experimental paradigms described in the literature for a wide variety of nCRF effects (e.g., end-stopping, surround suppression, contrast invariance of orientation tuning, cross-orientation suppression, etc.) to replicate these electrophysiology experiments on a dynamical system implementing optimal sparse coding. In the first contribution of this chapter, we show that this model produces individual units that reproduce

a wide variety of canonical nCRF effects. While another recent model [45] has also shown nearly all of these effects in a unified model along with some increased sparsity of the responses, the present work is the first to show that these effects can arise in a model that has only sparsity as the coding objective. In the second contribution of this chapter, when the population diversity of an nCRF effect has been reported in the literature (either through population statistics or multiple individual cells with varying response properties), we also show that this simulated population demonstrates much of the same population heterogeneity reported in the literature. Notably, the results we report are produced with a single set of model parameters (i.e., parameters are not tuned to produce each different effect), despite the system only being designed to optimize sparsity and not constructed to produce nCRF effects. These results show that the sparse coding hypothesis, when coupled with a biophysically plausible implementation, can provide a unified high-level functional interpretation to many population response properties that have generally been viewed through distinct models.

2.2 Results

2.2.1 Sparse coding and dynamical systems

The sparse coding model proposes that V1 encodes an image patch $I(x, y)$ with N pixels as approximately a linear superposition of M ($M > N$) dictionary elements $\{\phi_i(x, y)\}$,

$$I(x, y) \approx \sum_{i=1}^{M} a_i \phi_i(x, y), \tag{1}$$

where the coefficients $\{a_i\}$ represent the population activity (e.g., average firing rates) [12]. In this model, a neural population encoding the image $I(x, y)$ would calculate activity levels $\{a_i\}$ that minimize an energy function that is a weighted combination of a data fidelity term (e.g., mean-squared error) and a sparsity penalty (e.g., the coefficient magnitudes),

$$\sum_{x,y} \left(I(x, y) - \sum_{i=1}^{M} a_i \phi_i(x, y) \right)^2 + \lambda \sum_{i=1}^{M} |a_i|. \tag{2}$$

Here λ is a system parameter that controls the trade-off between the fidelity of the representation and the sparsity of the coefficients.

The sparse coding model is a functional model that can be implemented through many different mechanisms, including using generic convex optimization algorithms designed for digital computers. In this study we use a dynamical system proposed in [23] that employs neurally plausible computational primitives. Specifically, we implemented the sparse coding model by simulating the dynamical system given by:

$$\begin{aligned} \dot{u}_m(t) &= \frac{1}{\tau}\left[\langle \phi_m, \mathbf{I}(t)\rangle - u_m(t) - \sum_{i \neq m} \langle \phi_i, \phi_m \rangle a_i(t)\right] \\ a_m(t) &= T_\lambda(u_m(t)), \end{aligned} \tag{3}$$

where u_m is an internal state variable for each node (e.g., membrane potential), τ is the system time constant, and $\langle \cdot, \cdot \rangle$ is an inner product over the spatial dimensions. In the system dynamics, $\langle \phi_m, \mathbf{I}(t)\rangle$ captures the feedforward filtering while $\langle \phi_i, \phi_m \rangle a_i(t)$ captures the recurrent interactions that implement competition between cells to represent the stimulus. Note that the recurrent interaction between those cells is inhibitory if $\langle \phi_i, \phi_m \rangle > 0$ and excitatory if $\langle \phi_i, \phi_m \rangle < 0$ (since $a_i(t) > 0$ in our model). $T_\lambda(\cdot)$ is the soft thresholding function:

$$T_\lambda(u_m) = \begin{cases} u_m - \lambda & u_m > \lambda \\ 0 & |u_m| \leq \lambda \\ u_m + \lambda & u_m < -\lambda \end{cases}. \tag{4}$$

The input stimulus can be changed dynamically (e.g., a drifting sinusoidal grating), in which case the time-varying coefficients $\{a_i(t)\}$ will track approximate solutions, with the solution accuracy determined by the time scale of the input changes relative to the system dynamics. We note that recent theoretical work has demonstrated several network architectures that can efficiently implement other versions of sparse coding with various degrees of biological plausibility [38, 46, 47]. The network architecture being used in this study provably solves the optimization in Eq. (2) with strong convergence guarantees [48], can implement many variations of the sparse coding hypothesis (i.e., different sparsity-inducing cost functions) [49], and is implementable in neuromorphic analog circuits [50].

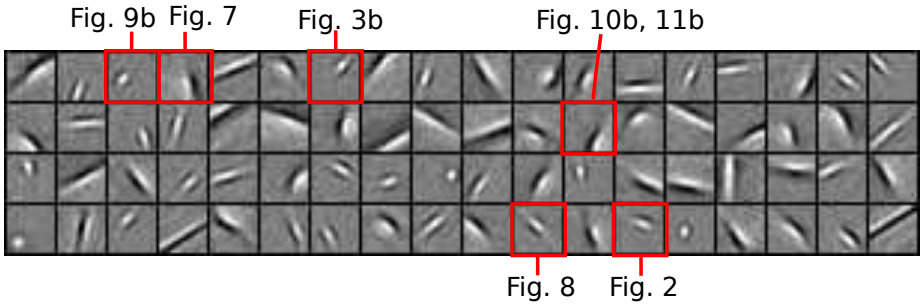


Figure 1: Subpopulation of dictionary elements ("CRFs") studied. The 72 dictionary elements that were recorded from in the model simulation. Dictionary elements were optimized for sparse coding under natural scenes (as described in the text) and selected for well-localized CRFs in the image patch. The units whose single cell activities are presented in later figures are indicated by red rectangles.

In our implementation, a dictionary $\{\phi_i(x, y)\}$ optimized for sparse coding with natural scenes was determined via unsupervised learning under sparsity constraints using whitened natural scenes as the training set (whitening is a first-order approximation of retinal processing). The learned dictionary was overcomplete with $M = 1024$ effective dictionary elements for the 16×16 pixel image patches used as stimuli. The training set, whitening and learning rule were all exactly as in [12], while the sparse codes during training (i.e., solutions to (2)) were calculated using a standard software package [51] (for computational efficiency) with $\lambda = 0.6$. We interpret these dictionary elements as the classical spatial receptive fields (CRFs) of the simulated neurons. This interpretation is supported by our own simulated receptive field mapping experiment (results not shown) using sparse dot stimuli, similar to previous studies (e.g., see Fig.4b in [12]). The results demonstrated in this study are based on the responses of 72 units in this dictionary that had CRFs well-localized within the available image patch (shown in Fig. 1).

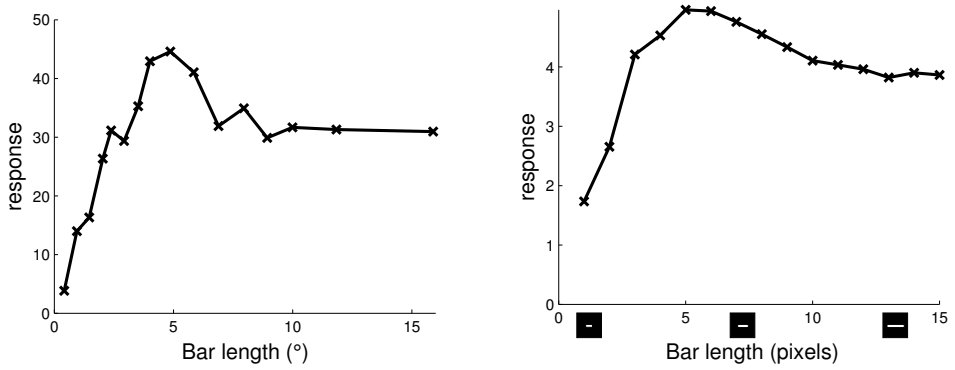
The system parameters described above (i.e., membrane time constant, sparsity level λ) are kept the same for every simulation in this chapter (details given in Materials and Methods). In other words, no attempt was made to tailor the system to reproduce each effect individually (some interesting exceptions where parameter changes correspond to apparently conflicting results in the literature are shown in Sect. 6.1). We interpret the

sparse coefficients a_m in Eq. (2) as the trial-averaged instantaneous spike rate of neurons in the model population. To do this, we also impose a positivity constraint $a_m \geq 0$ and extend the dictionary matrix by including both the original dictionary elements and the negative of the dictionary elements (i.e., doubling the size of the matrix to use the same effective dictionary as if there were both positive and negative coefficients). This mirrored receptive field structure is reminiscent of the push-pull feedforward input structure in the visual simple cells [52].

In the following sections, we highlight several common nCRF effects from the literature and illustrate that this sparse coding model can largely reproduce both reported individual response properties and much of the reported response diversity across V1 neurons. For each nCRF effect the simulation was constructed to match as closely as possible the experimental protocol described in the experimental procedures section of the corresponding electrophysiology paper, including stimulus construction parameters and data analysis (details given in Materials and Methods). We classify the studied nCRF effects into three groups: suppressive effects that are evoked by the presence of stimuli outside the classical receptive field (CRF surround effects), effects where the response modulation depends on the orientation of the stimulus in the surround (CRF surround orientation effects) and effects that reflect the nonlinearity of the CRF center (nonlinear CRF effects).

2.2.2 CRF surround effects

Stimuli in the region surrounding the CRF can have a modulatory effect on a neuron's response despite not inducing significant response in isolation (by definition of the CRF). In perhaps the simplest form of this suppressive modulation, it has long been known that some V1 neurons exhibit *end-stopping* where the spike rate decreases for a cell responding to an optimally-oriented bar stimulus when the bar length is increased beyond the CRF boundaries. An example figure depicting the end-stopping effect as observed in cat electrophysiology recordings [1] is reproduced in Fig. 2a. When simulating this experiment [1] on the sparse coding model, some of the model cells (such as the target cell shown in Fig. 2b)

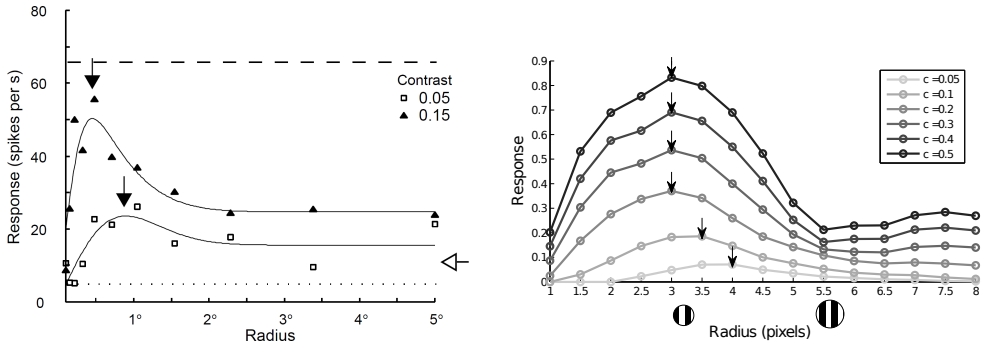


(a) (b)

Figure 2: End-stopping. (a) End-stopping response in a simple cell from cat V1 responding to an optimally-oriented light bar stimulus (data replotted from [1, Figure 1]). (b) The length tuning curve of a simulated sparse coding model neuron (target) demonstrates end-stopping behavior.

exhibit the same characteristic suppression with increasing bar length. The end-stopping effect was previously shown in [53] to emerge in the sparse coding model. The end-stopping effect can be simply understood in terms of the goals of sparse coding. When the bar is short, the CRF of the target cell is the most efficient description of the stimulus and that cell has the strongest response. However, when the bar is long enough that it is better explained by the CRFs of other cells, the target cell becomes suppressed by these competitors so as to maintain a sparse representation. The Discussion section contains a detailed look at how the network interactions supporting the sparse coding model can produce this effect.

Similar to end-stopping, some V1 neurons also exhibit *surround suppression* where their response to a sinusoidal grating patch decreases as the patch size increases beyond the CRF. Additionally, the tuning curve for patch size often exhibits *receptive field expansion* at low contrast, meaning that the patch size achieving the maximum response increases at low contrast (Fig. 3a). As illustrated in the response of an example model cell shown in Fig. 3b, the sparse coding model can exhibit the same basic suppression and receptive field expansion properties observed in electrophysiology experiments. In addition, we note that the slight increase of response level (i.e., response rebound) at large stimulus size visible in Fig. 3b is also visible in Fig. 3a and discussed elsewhere [54].

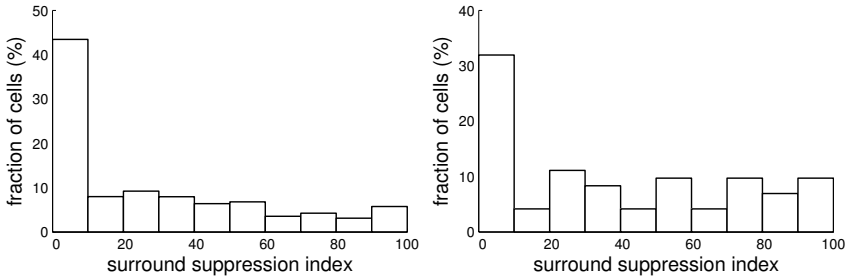


(a) (b)

Figure 3: Surround suppression and RF expansion in a single cell. (a) A plot illustrating that cortical neurons show surround suppression and expansion of CRF size at low contrast (reprinted by permission from Macmillan Publishers Ltd: Nature Neuroscience, Figure 1a from [2]). (b) The size tuning curve of a simulated sparse coding model neuron at various contrast levels ("c" stands for contrast, with lighter curves representing lower contrast). The model neuron exhibits two characteristic behaviors reported in the electrophysiology literature: suppression with increasing stimulus size and an increase in the optimal stimulus size with lower contrast. The maximum of each tuning curve is marked by an arrow.

The network interactions giving rise to surround suppression are presumably similar to that of end-stopping, but are more difficult to specify given the added dynamics of the drifting grating stimulus. In particular, due to the suboptimal match of the target CRF to the larger stimulus, competition from other cells (that better match the larger stimulus) can suppress the target cell's response. This competition can also be modulated by the stimulus contrast and may contribute to the receptive field expansion. Specifically, at low contrast the competing cells have lower response levels (resulting in a weaker suppressive effect on the target cell), enabling the response of the target cell to grow with the stimulus size.

Despite the evidence detailed above that some biological and model V1 neurons exhibit surround suppression, a single example cell is insufficient to quantify the prevalence of this effect in a population encoding sensory information. While many nCRF effects are reported as single cell response properties, some studies have attempted to quantify how strongly an effect is expressed across the population. In the case of surround suppression, two metrics have been used to quantify the degree of suppression and receptive field expansion demonstrated by a cell. One is the suppression index (SI), calculated as the ratio between the

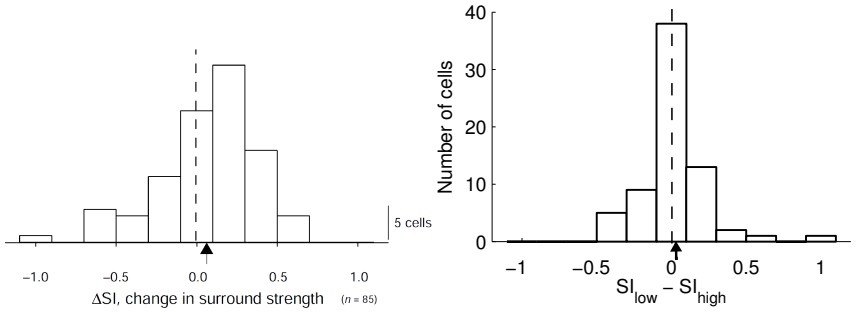


(a) (b)

Figure 4: Surround suppression index distribution. (a) Physiologically measured distribution of surround suppression index (SI) in cat V1 (data replotted from [3, Figure 2A]), illustrating that most cells do not exhibit significant surround suppression and the SI distribution is relatively uniform among suppressive cells. (b) The SI distribution for the model cells, illustrating the same qualitative properties as the distribution in (a).

(suppressed) response value at large stimulus sizes and the peak response value (indicated by arrows in Fig. 3b). The second metric is the RF expansion ratio, calculated as the ratio of the size tuning peak location at high contrast against that at low contrast.

In many physiological studies (both in monkeys [5] and in cats [3]), a large proportion of cells actually show little suppression, with relatively few cells exhibiting strong suppression. An example SI distribution from cat V1 is shown in Fig. 4a, demonstrating a dominant peak at zero suppression and a relatively uniform distribution among more suppressive cells. A similar population distribution emerges from the sparse coding model cells, as illustrated in Fig. 4b. Another characteristic of the surround suppression index is that it is largely invariant to the stimulus contrast. In other words, the difference in SI at high and low contrast is close to zero (Fig. 5a) with a mean value of 0.06 [2]. We also observed this characteristic in the sparse coding model cells (Fig. 5b), with a mean SI difference of 0.02. We note here that some studies (e.g. [15]) recorded unusually high percentage of cells showing significant surround suppression, perhaps due to a different experimental preparation. Interestingly, the sparse coding model can qualitatively reproduce these apparently conflicting results by using a different set of parameters to encourage more sparsity (see Fig. 39).



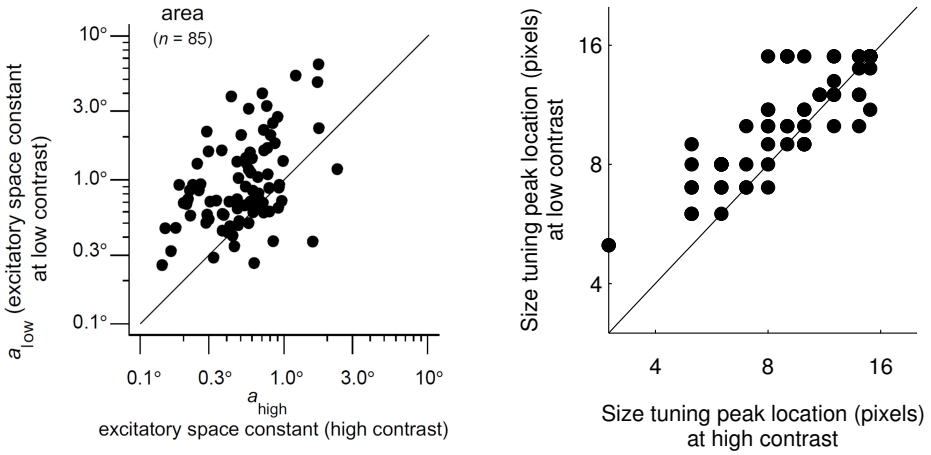
(a) (b)

Figure 5: Distribution of the SI difference. (a) Distribution of the SI difference (ΔSI) between low and high contrast levels in macaque V1 (reprinted by permission from Macmillan Publishers Ltd: Nature Neuroscience, Figure 6b from [2]). The mean difference is 0.06, demonstrating that on average the SI for a cell is contrast invariant. (b) The distribution of ΔSI for the sparse coding model cells. The mean difference is 0.02, also demonstrating contrast invariance in SI.

A scatterplot of RF expansion ratios for V1 cells in macaque [2] shows clearly that on average, the CRF size is larger at low contrast than at high contrast (Fig. 6a). A scatterplot of expansion ratios for the sparse coding model population shows the same qualitative trend of expanding CRF size at low contrast. We note that the mean expansion ratio in the sparse coding model cells (1.16) is lower than typically reported values in the electrophysiology literature (e.g., 2.3 in [2]). This quantitative difference may be due to variations in the RF expansion ratio definitions (e.g., the study in [2] uses a difference of Gaussians fit rather than tuning curve peaks), the lack of contrast saturation in the present model (see Discussions), or biased sampling of neurons in the electrophysiology literature [17]. The possibility that the true expansion ratio might be lower than previously reported is corroborated by a recent study reporting that as many as 40% of cat V1 neurons show length tuning peaks that are invariant to contrast changes [55].

2.2.3 CRF surround orientation effects

The modulatory effects seen from surround stimulation can depend on a number of stimulus properties, including contrast, spatial extent (relative to the CRF), and stimulus orientation in the surround. In particular, modulation is often most suppressive when the surrounding



(a) (b)

Figure 6: Size tuning peak at high vs. low contrast. (a) RF expansion of macaque V1 cells (reprinted by permission from Macmillan Publishers Ltd: Nature Neuroscience, Figure 3a from [2]). (b) RF expansion of sparse coding model cells. Most points lie above the diagonal, indicating that (on average) the optimal stimulus size is larger at lower contrasts and the cell demonstrates RF expansion.

stimuli are at orientations parallel to the preferred CRF orientation (iso-oriented), and less suppressive (or even facilitatory) when the stimuli are perpendicular to the preferred CRF orientation (ortho-oriented). For example, when stimulating a cell with an optimally oriented sinusoidal grating just covering the CRF (i.e., the orientation eliciting the strongest response), a grating in the annulus surrounding the CRF often suppresses the cell when it is iso-oriented and has little effect when it is ortho-oriented. An example of this *surround orientation tuning* in macaque V1 cells [4] is shown in Fig. 7a. The sparse coding model cells can also demonstrate the same type of surround orientation tuning, as illustrated by the model cell response shown in Fig. 7b. This tuning behavior in the model is likely due to the difference in the strength of competition with different stimulus surround orientations. In particular, the competing cells stimulated by iso-oriented surrounds are likely to have stronger CRF overlaps with the target cell and therefore induce more competition than the cells stimulated by ortho-oriented surround stimuli.

Orientation tuned surround effects can have substantial variations, even with minor changes in the stimulus. For example, the modulatory effect can be facilitatory at some

surround orientations, causing a net increase in the response of the cell to CRF stimulation alone. This facilitatory effect is often seen when using a center stimulus slightly larger than the optimal size [4], as shown in Fig. 7c for the same cell as in Fig. 7a. Interestingly, increasing the size of the center stimulus for a model cell can likewise induce facilitation when the surround stimulus is close to ortho-oriented (shown in Fig. 7d for the same cell as in Fig. 7b).

As with surround suppression, a single example of facilitation in the surround orientation tuning does not characterize the prevalence of this effect in a population of V1 cells encoding a stimulus. The degree of facilitation expressed by a neuron can be characterized by measuring the ratio between the maximum of the surround orientation tuning (the maximum of the solid line in Fig. 7b) and the response to the center at the optimal orientation with no surrounding stimulus (the maximum of the dashed line in Fig. 7b). In macaque V1 [56], the median of the facilitation ratio across the measured population was found to be 1.44 at high contrast and 1.71 at low contrast. The sparse coding model cells show a similar dependency on contrast levels, with the median facilitation ratio ranging from 1.15 at high contrast and 1.31 at low contrast.

The surround orientation tuning properties described above can be substantially influenced by the contrast difference between the center and the surround. For example, if the center contrast is fixed and the surround contrast is varied, the most significant suppression in individual macaque neurons was observed with the iso-orientated stimuli at high surround contrast (see Fig. 8a) [5]. Similarly, when plotting the responses as a function of center contrast for various surround settings (e.g., no surround, iso-oriented, and ortho-oriented), the suppressive effects in macaque were most pronounced with the iso-oriented stimuli at high center contrast (see Fig. 8c) [6]. Both of these dependencies on contrast can also be observed in the sparse coding model cells, as shown in Fig. 8d and Fig. 8b. Again we note that in some physiological studies an apparently conflicting result is reported where cat V1 neurons show facilitation with iso-oriented surround stimuli at low

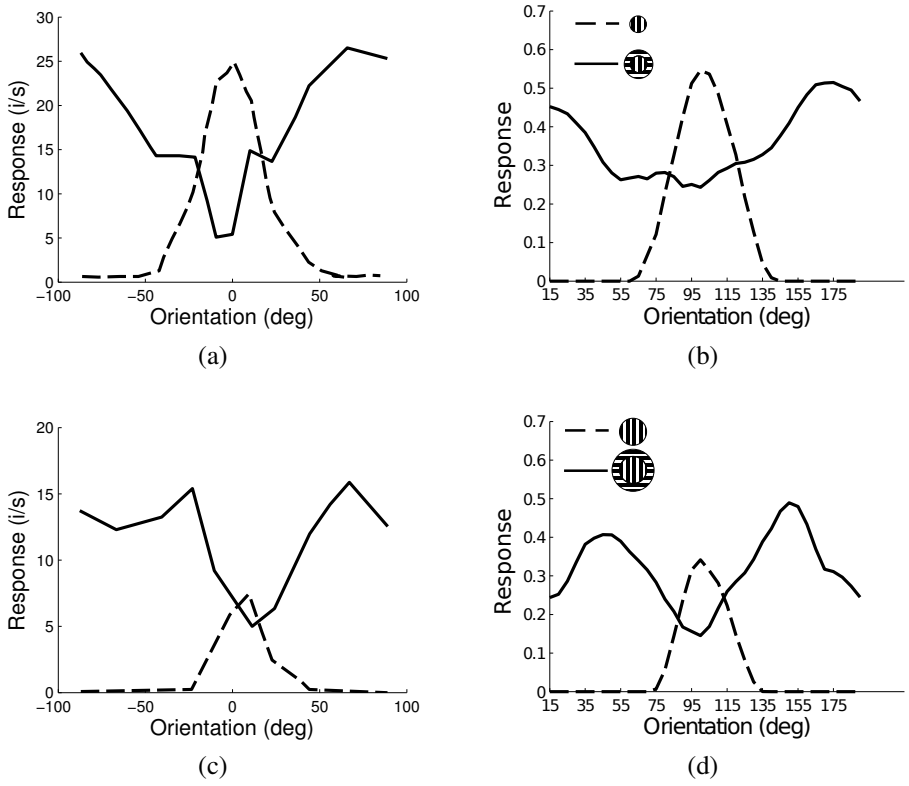


Figure 7: Orientation tunings for surround suppression and facilitation. (a) Center and surround tunings with the optimal stimulus center size in macaque V1 (data replotted from [4, Figure 2A]). The center orientation tuning curve (dashed line) shows the cell's response to a CRF sinusoidal grating. With the CRF stimulus fixed to an optimally-oriented grating, the surround orientation tuning curve (solid line) shows the cell's response to a sinusoidal grating in the annular surround at various orientations. (b) A sparse coding model cell demonstrating similar surround orientation tuning properties, with highest levels of suppression at iso-oriented surround stimuli and almost no suppression for ortho-oriented surround stimuli. (c) Center and surround orientation tunings of the same cell as in (a) with the stimulus center size increased beyond the CRF and the width of the surround annulus unchanged (data replotted from [4, Figure 2B]). (d) The same sparse coding model cell as in (b) demonstrates the facilitatory effects at ortho-oriented stimuli seen in (c).

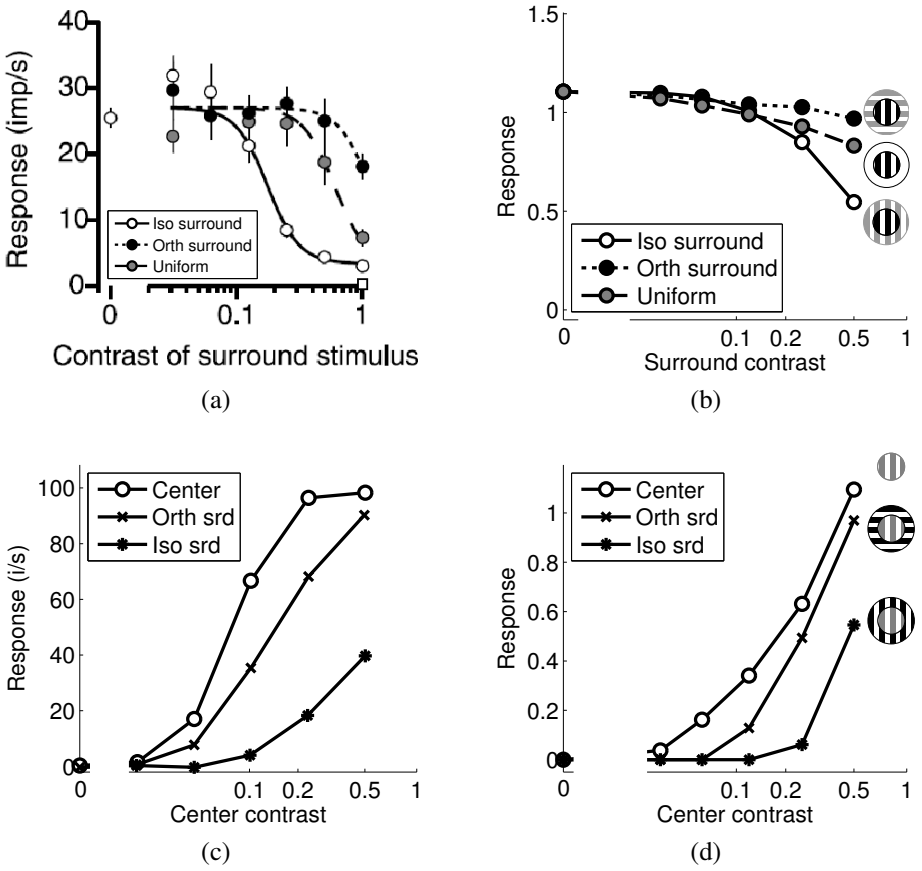


Figure 8: The effect of contrast on surround influences. (a) Surround contrast tunings with fixed center contrast in macaque V1 and varying surround stimuli (reprinted by permission from the Society for Neuroscience: The Journal of Neuroscience, Figure 6B from [5]). The gray markers correspond to responses to a uniform surround at different contrast. (b) Surround contrast tunings with fixed center contrast in the sparse coding model. As with the neuron responses in (a), the model cell is most suppressed for iso-oriented surround stimuli at high contrast. (c) Center contrast tunings with fixed surround contrast in macaque V1 simple cells with varying surround orientations (data replotted from [6, Figure 5A]). (d) Center contrast tunings with fixed surround contrast in the sparse coding model. As with the neuron responses in (c), the model cell shows that (especially at high contrast) an iso-oriented surround (asterisk markers) is more effective than an orthogonal surround (cross markers) at suppressing the response to the center alone (white circle markers). As mentioned in the text (see Discussions), the lack of contrast saturation in the present sparse coding model is evident in this figure by the model response at high contrast.

CRF contrast [16] (Fig. 40a). Interestingly, the sparse coding model can also reproduce this behavior when using a different set of parameters (see Fig. 40b).

2.2.4 Nonlinear CRF effects

Even when the stimulation is confined to the CRF with no involvement of the surround, cells in V1 exhibit several nonlinear effects that cannot be explained by a canonical linear-nonlinear model [28]. One example of such an effect is the *contrast invariance of orientation tuning* for V1 cells. In a linear-nonlinear model based on CRFs, higher contrast stimuli evoke stronger responses that more readily exceed the spiking threshold, thus broadening the orientation tuning curve for higher contrast stimuli (the "iceberg effect" [57]). However, as reported in the cat physiology literature, the orientation tuning width is largely contrast invariant [7] as demonstrated in Fig. 9a. Cells from the sparse coding model can also display this contrast invariance in the width of their orientation tuning curves, as shown in Fig. 9b. This invariance can potentially be attributed to recurrent inhibition from competing cells at orientations where the target cell is not the most efficient description (e.g., ortho-oriented stimuli). Even though these competing cells may not have large overlap with the CRF of the target cell, as the contrast increases they will become more active and induce stronger inhibition, thereby narrowing the tuning width of the target cell compared to the low-contrast response. Indeed, compared to the predictions of a linear-nonlinear model (not shown), the tuning width from our model is much narrower.

The degree to which the width of the orientation tuning curve changes for a cell can be quantitatively measured by calculating the half-width at half-height of the Gaussian fit to the tuning curve for various contrast levels [8]. The population statistics can be plotted as a histogram tabulating the slope of the best linear fit to the width expansion with contrast for each cell. An example of this measure from ferret V1 demonstrating that the tuning curve width is almost constant with contrast is shown in Fig. 9c [8]. In this same measure, the sparse coding model also exhibits strong contrast invariance properties across the population, as shown in Fig. 9d. Both the ferret V1 population and the sparse coding model have

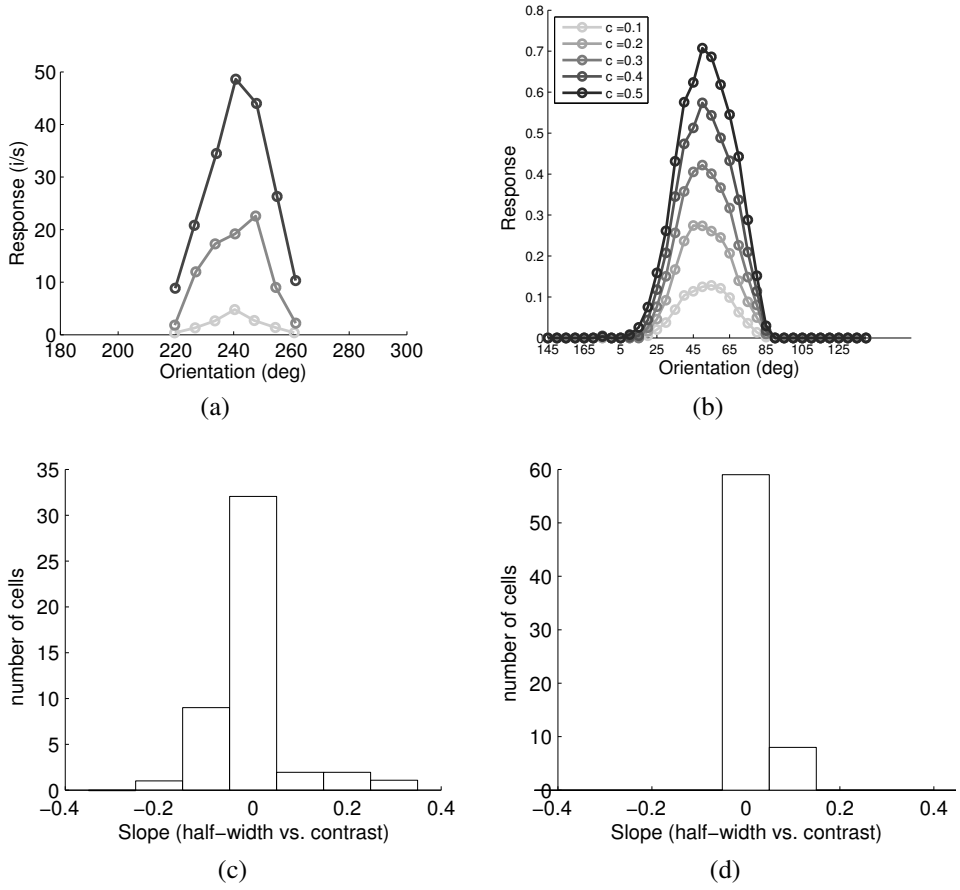


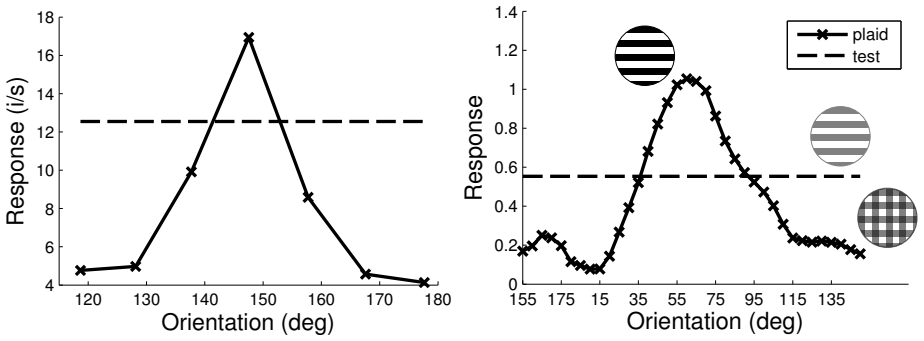
Figure 9: Contrast invariant orientation tuning. (a) Contrast invariance of orientation tuning curves recorded in cat V1 (data replotted from [7, Figure 3A]). Note that the width of the orientation tuning curve does not change with contrast. (b) Sparse coding model neuron that demonstrates the same invariance property. Lighter curves correspond to lower contrast ("c" denotes contrast level). (c) Distribution of the slope of tuning curve half-width vs. the contrast in ferret V1 (data replotted from [8, Figure 3B]). The sharp distribution around 0 indicates that the tuning curve half-width is contrast invariant (mean value is 0.002). (d) Distribution of the half-width vs. the contrast slope in the sparse coding model cells (mean value is 0.032). The model cells clearly demonstrate contrast invariance of the tuning curve half-width, and an even tighter peak around zero slope than shown in (c).

a slope tightly concentrated around zero in these histograms, with mean values of 0.002 and 0.032 respectively. The mean values of the half-width at high contrast measured in physiology (16.1 ± 1.1°) [8] and the model (13.87 ± 5.84°) are also similar.

An example of a nonlinear CRF effect using a more complex stimulus is *cross orientation suppression*, where a plaid (i.e., an ortho-oriented mask grating superimposed on an iso-oriented test grating) suppresses the response of the cell to the test alone. Fig. 10a and Fig. 10b show examples of this suppressive tuning property from cat V1 [9], as well as from a single cell in the sparse coding model. This kind of facilitatory effect may be due to a number of factors, including excitatory connections between cells (i.e., other cells in the population encouraging the target cell to represent the stimulus when they are unable to do so) or dis-inhibition, where a distant cell inhibits an intermediate cell that subsequently releases an inhibitory effect on the target cell [58].

The degree of cross orientation suppression depends on other factors beyond the orientation of the mask stimulus, including the contrast levels of the test stimulus. This contrast dependency was observed in cat V1 (shown in Fig. 11a) [10], and is also visible in the sparse coding model neurons as shown in Fig. 11b. Note that while the qualitative trends in the contrast dependency are the same in the model and in physiology, the lack of contrast saturation in the present model is evident in this figure (see Discussions).

The degree of cross orientation suppression expressed in a population of cells can be characterized by comparing the response to the plaid with the response to the test alone. A scatter plot of the normalized spike rate of cat V1 cells shown in Fig. 12a for the test versus plaid stimuli demonstrates that most cells have a suppressive response to the plaid (as depicted in the single cell response in Fig. 10a) [11]. Furthermore, the scatter plot indicates that the suppression is more pronounced for lower test contrasts. As shown in Fig. 12b, the sparse coding model population exhibits the same qualitative properties, with most cells exhibiting plaid suppression that increases with lower test contrast. Quantitatively, the mean cross orientation suppression ratio between the test and plaid responses for cat



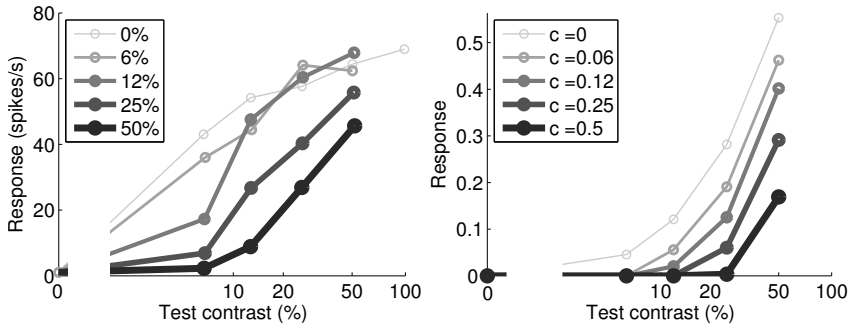
(a) (b)

Figure 10: Cross orientation suppression. (a) A cat V1 simple cell demonstrates cross orientation suppression by responding with lower firing levels to an iso-oriented test grating if an ortho-oriented grating (mask) is superimposed (data replotted from [9, Figure 3(A)]). The dashed line is the response to the iso-oriented test grating with no mask stimulus. (b) Cross orientation suppression exhibited by a sparse coding model neuron. Note the same dependence on the orientation of the mask that is seen in (a).

V1 was measured at 0.11 for low test contrast and 0.71 for high test contrast [11]. The sparse coding model cells have mean cross orientation suppression ratios of 0.59 and 0.95 for low and high test contrasts (respectively). While the model shows the same qualitative trend and overlaps in range, the specific values for these ratios are slightly higher than the reported experimental values. This small quantitative discrepancy might be due to the presence of contrast saturation in the physiology (visible in Fig. 11a) and its absence in the sparse coding model (Fig. 11b; see Discussions).

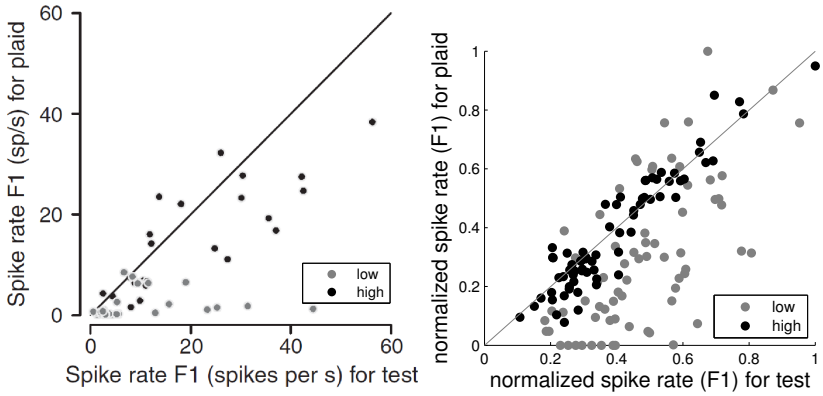
2.3 Discussion

Electrophysiology research in V1 has revealed a wide variety of nCRF effects that may appear to be due to many different aspects of neural coding or cortical processing. The functional interpretation of these effects is especially complex given the heterogeneity of the responses exhibited across populations of cells reported in the literature. We have demonstrated for a wide variety of nCRF effects that both the canonical individual cell response properties and a substantial diversity of population response properties are emergent characteristics of a simple dynamical system implementing a sparse coding model. This



(a) (b)

Figure 11: Contrast tuning of the plaid. (a) Contrast tuning curves of the test at different fixed mask contrast levels for a cat simple cell (data replotted from [10, Figure 2A]). (b) Contrast tuning curves of the test for the same sparse coding model cell as in 10b. Note again the same response modulation as in physiology despite the lack of contrast saturation in the model (see Discussions).



(a) (b)

Figure 12: Population distribution of cross orientation suppression. (a) Measurement of modulation (F1) component of the response to a test grating alone vs. that with a superimposed orthogonal grating from a population of visual cortical neurons in cat (reprinted by permission from Macmillan Publishers Ltd: Nature Neuroscience, Figure 2b from [11]). The unity line represents where there is no suppression. The response at low test contrast is further away from the diagonal, suggesting more suppression in this regime. (b) Measurement of F1 response to a test grating alone vs. that with a superimposed orthogonal grating from the sparse coding model population. Note that just as in the physiology data, the model has the same general suppressive behavior, with increased suppression with lower test contrast.

model appears to produce a very good qualitative match to many measures of population response statistics, and in many cases produces quantitative measures of these statistics that are in a similar range to reports in the physiology literature. By demonstrating a coding model that can account for these response properties, these results provide a potential functional insight into the role of nCRF effects in optimal sensory coding. While not mutually exclusive of other functional models that may also play a role in neural coding, the sparse coding model is one of the few models (along with [59]) able to substantially reproduce some nCRF effects as well as account for the emergence of localized, oriented, and frequency-selective CRFs [12]. In particular, despite not being constructed to produce nCRF effects, the present model appears able to capture population properties of nCRF effects that have been difficult for other functional models to produce (e.g. the contrast invariance of surround suppression index in Fig. 5b, as discussed in [60]).

There are several existing results that share a similar goal of providing high-level functional interpretation of nCRF effects. Perhaps most closely related to the present study is the PC/BC model [45, 59, 61–63], which has also been able to reproduce most of the nCRF effects demonstrated in this paper [45]. It is interesting to note that although it has other functional goals, the PC/BC model does exhibit high sparsity [59] and has accounted for classic CRF tuning properties [59]. While there is significant overlap in the demonstrated nCRF effects, the present work is unique in exhibiting the sufficiency of a model derived from sparse coding to produce the observed effects and to reproduce the population diversity seen in physiology (which the PC/BC model has yet to demonstrate). Given the similar behavior of the PC/BC model and the present model, it is possible that there is a deeper underlying relationship between the PC/BC model and sparse coding than is presently understood. Other example related works include the basic predictive coding model [64], where a subpopulation of model neurons communicating prediction errors exhibits some of the single cell nCRF effects documented in the present study. Another example is the

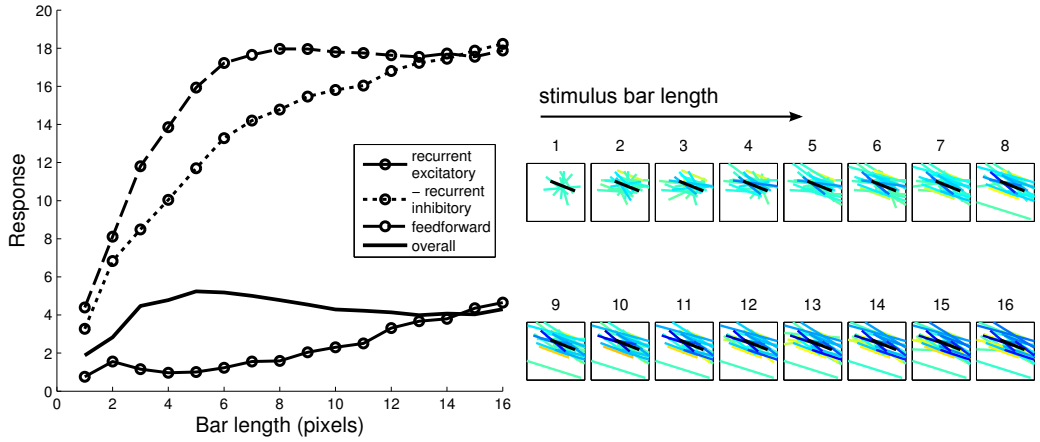
divisive normalization model [65], where contextual effects emerge from a population interaction that modulates the gain in an attempt to maximize the independence of neighboring units. While both of these models account for some individual effects, they are not currently known to reproduce the population diversity seen in physiology or to alone be sufficient to also account for the emergence of known CRF properties (without an added sparsity constraint). More recent models capture the center-surround homogeneity (e.g. orientation co-alignment) in the natural scenes through a generalized form of divisive normalization [66] or capture the covariance structure between pixels in natural scenes [67]. While each of these models demonstrates some individual nCRF effects, these models are also not currently known to reproduce the population diversity seen in physiology (in particular, [66] simulates responses using a single generic unit and not a diverse population) and neither model currently has a fully specified implementation in a biologically plausible circuit (although an approximate form of the model in [67] may enable such an implementation). Another related model was described in [68], which demonstrated that a spiking input targeted divisive inhibition mechanism gives rise to competition among sensory feature detectors and non-classical-like effects. While this model have some interesting features that the present model does not have (e.g., biologically realistic spiking behavior), the stimuli and CRF representations were 1D idealized functions and it's not clear how the results extend to 2D images.

An important feature of the present work is that the same model (with the same parameters) is used to produce all of the presented results (i.e., parameters were tuned once and fixed for all experiments in the main text). The qualitative and quantitative matches observed in this chapter rely on these parameter settings combined with the dynamical system implementation of the sparse coding rule. For example, changes in the system that would actually encourage responses with higher sparsity (e.g., increasing λ, solving Eq. (2) using a conventional digital algorithm, running the dynamical system implementation with more integration time steps/faster non-biological time constants) would often generate similar

single cell nCRF effects [69] as presented here (results not shown), but those effects would be too strong to be a quantitative match to the population properties (e.g., a far higher percentage of model cells would show strong surround suppression than is reported in physiology; see Fig. 38). Sect. 6.1 demonstrates some instances where simple parameter changes in the model can actually account for apparently conflicting reports regarding nCRF effects in the experimental literature. We speculate that different settings of λ in the model may reflect differences in experimental preparations, such as different species and various levels of anesthesia. Indeed, anesthesia is known to influence the sparsity level in sensory systems [70, 71], and some perceptual contextual effects only occur in awake animals [72]. These observations about changes in the results with varying sparsity levels indicates that the sparse coding objective appears to be sufficient to produce the nCRF modulations, but the dynamical system implementation (with biophysical time constants) is required to produce the heterogeneity necessary to be a good quantitative fit. We also note that the role the dynamical system plays in the present work is similar to recent work [38] showing that learned dictionaries can be a much better quantitative match with measured macaque CRFs when the sparse coding model is implemented in a neurally-plausible network model. It is presently unclear if a different dynamical system minimizing the sparse coding objective would also result in the heterogeneity necessary to still be a good quantitative fit to physiology. Similar variations in the quantitative fits (especially to population data) are expected when using other sparsity penalties beyond the ℓ_1 norm used here [49], or when using sparse coding implementations that encourage more "hard" sparsity (i.e., more elements that are exactly zero) [38]. In a similar vein, the present study uses a four-times overcomplete dictionary optimized for sparsity under natural scenes, and this model component is also likely important to the presented results. Though investigating the role of the dictionary would be an interesting avenue of further exploration, we expect that larger dictionaries may enable more sparse responses which also may demonstrate more suppression than what is seen in the current model.

The recurrent interactions between cells in the sparse coding model implement a rich nonlinear response where cells compete to represent stimulus features. While it has been noted that stimuli in the CRF surround can produce sparse responses [29, 30, 32], the surprising finding of this work is that the particular form of inhibition and excitation necessary to implement a sparse coding model is sufficient to explain so many individual and population nCRF properties. At a high level, these effects likely arise from the present model because the observed responses produce a more efficient representation of the stimulus than alternative population responses. While a detailed investigation of how the network interactions give rise to the response properties is an interesting open question for future investigation, in general this is difficult to determine due to the interactions between the network dynamics and the stimulus dynamics (i.e., the response properties arise from the average response over a drifting grating, in addition to being influenced by network dynamics). In the case of end-stopping, the stimuli is not drifting and we can see more explicitly how this effect arises from the principles of sparse coding. In response to a given fixed stimuli, the steady-state network response is composed of a combination of feedforward excitation, recurrent excitation and recurrent inhibition. When plotting these three components of the steady-state response as a function of the bar length (Fig. 13a), it is evident that the overall response is mostly driven by the feedforward component and the recurrent inhibition. The feedforward excitation saturates as a result of the stimulus growing out of the CRF, but the recurrent inhibition keeps growing with increased bar length. To see the spatial extent of the recurrent influence, Fig. 13b shows the CRF locations and orientations of the cells influencing the target cell. As expected, inhibition mostly comes from cells with overlapping and co-linear CRFs that represent a more efficient description of the stimulus as the bar length increases.

There has been a long history of debate over the mechanisms underlying various nCRF effects [27], with each effect generally having a substantial literature attempting to answer questions about the detailed aspects underlying the modulatory response properties



(a) (b)

Figure 13: Decomposition of the recurrent inputs contributing to the end-stopping effect. (a) Overall decomposition of the response into recurrent excitatory, inhibitory, and feedforward components; (b) Locations and orientations of the CRFs of cells contributing to the recurrent excitatory and inhibitory signals at different bar lengths. Only CRFs with significant influences are displayed (i.e., $|\langle \phi_i, \phi_m \rangle| a_i(t) > 0.1$ at steady state). The warmer color (yellow) represents the location and orientation of the CRFs for cells contributing to recurrent excitation, the cooler color (blue and cyan) represents the CRFs for cells contributing to recurrent inhibition. Higher contrast in the color indicates a stronger excitatory or inhibitory effect on the target cell. The black bar represents the target cell CRF. Note that as the bar length increases, the suppressive effect is mostly due to recurrent inhibition from cells that are a better description of the new stimulus (and therefore would be a more efficient stimulus description according to the sparse coding model).

(e.g., the relative role of intra-cortical connections versus feedforward projections from thalamus in contrast invariant orientation tuning [73], as well as the role of feedback connections [74]). The implementation used in this work (see Materials and Methods) would appear to suggest that these contextual effects can emerge from recurrent network structure in the absence of nonlinearities in the thalamic input or feedback from higher cortical areas. However, mechanistic interpretation of functional models must be cautious as there are often many possible mappings of the model to circuitry and biophysical mechanisms. For example, past work has shown that it is possible to have mappings of functional models onto circuitry that are very different from their original intuitive mappings (e.g., divisive normalization [75] and predictive coding [61]). The sparse coding dynamical system used in this study is open to the same variety of mechanistic interpretations. For example, the recurrent inhibitory influences could be implemented [76] via local inhibitory interneurons receiving convergent inputs from local excitatory neurons [77] and having dense (many-to-one) output connections with these excitatory neurons [78]. Alternately, it is possible that these inhibitory influences could be implemented via a mechanism based on long term depression of synaptic connections between excitatory cells in cortical layer 4 [79] and global inhibition [80]. For another example, as demonstrated in [75], it might be possible to achieve similar computational goals through nonlinearities in the feed-forward thalamocortical circuit, rather than a recurrent network. For yet another example, the recurrent competition could be implemented through subtraction as in our model, or through division as in [45]. It remains an open question to determine the most biophysically appropriate mapping of the present model onto a circuit implementation.

While the mechanisms underlying individual nCRF effects is an interesting area of investigation, another related question of interest is to determine which aspects of the model are responsible for the observed population variability. In the present model, the dictionary serves to define both the activity driving each cell through the CRF, as well as determining the synaptic weights that define the recurrent influences in the network dynamics. Because

the present dictionary was learned from the sparse coding objective on natural images, it is optimal for this coding strategy and demonstrates significant variability as observed in biological CRFs. While a detailed investigation of how the model gives rise to the response diversity is also a challenging and interesting open question for future investigation, one interesting preliminary question is what role the variability in the dictionary plays in the observed nCRF response variability. As a specific example, we have found the surround suppression index to be significantly anti-correlated with the CRF size (Fig. 14; correlation coefficient = -0.89; $p < 0.001$). While we are unaware of studies investigating this relation in the physiology literature, there are several studies that do suggest this type of anti-correlation. One piece of evidence [81] shows that cortical layers with larger CRFs also tend to have lower SIs and vice versa. Another corroborating study [82] shows that suppressive V1 cells have smaller CRFs compared to plateaued and facilitative cells. This anti-correlation may be present simply because there are fewer cells with larger CRF size in the model (visible in Fig. 14) and in V1 [83], making these cells more likely to be used in an efficient coding model whenever the stimulus grows past a certain size. It is also possible that the limited stimulus sizes used in the current model and many physiology studies (e.g. [84]) could be producing a boundary effect that contributes to some of these observations. It is presently unclear if the inherent variability in the dictionary is alone sufficient to produce the response variability observed in biology (i.e., if another coding model could produce this same variability when using CRFs from this same type of learned dictionary) or if significant response heterogeneity requires the interaction of a learned dictionary with a dynamical system implementing sparse coding.

Some contextual effects, especially ones that involve perception such as perceptual pop-out, figure ground segregation [37], and contour integration [36] operate over a larger range (e.g. over 8 times the CRF size in [85]) and are likely to be mediated by long-range lateral connections [86]. The present study did not test the emergence of these types of effects in the sparse coding model due to the limited size of the dictionary elements. The sparse

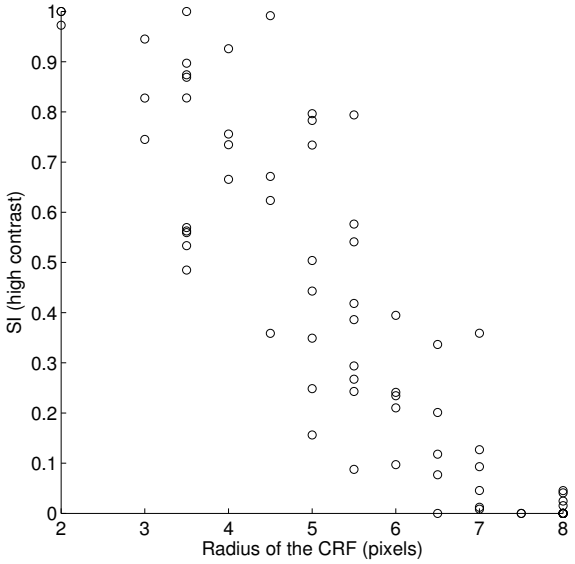


Figure 14: Surround suppression index is anti-correlated with the CRF size. Cells with larger CRFs tend to be less suppressed by a surround stimulus (correlation coefficient = -0.89; $p < 0.001$). The level of suppression is measured by the suppression index (SI) at high stimulus contrast.

coding model simulated here used a substantially overcomplete dictionary (see Materials and Methods), thus the size of the visual field we were able to simulate is limited by the current computational complexity of learning large scale dictionaries from the statistics of natural images. While it may seem unlikely that long-range effects could emerge from the present model when the only direct influences are between cells with overlapping receptive fields (see Materials and Methods), it is conceivable that second order effects (e.g., dis-inhibition, where a distant cell inhibits an intermediate cell that subsequently releases an inhibitory effect on the target cell) may play a central role that would only be discovered in a study using larger visual fields. An alternative is to incorporate long-range lateral connections explicitly into a sparse coding model [87].

Despite the wide variety of nonlinear properties observed in the sparse coding model, this model alone is unable to reproduce some nCRF effects because it lacks the stereotypical saturating contrast response function [88]. While this contrast saturation would be a simple addition to the model, the present study focuses on the basic sparse coding model

to isolate the response properties due to the nonlinear interactions required to achieve sparsity. It is interesting to note that the model can still reproduce several contrast dependent contextual effects even without an explicit contrast saturation mechanism. Indeed, it has been previously suggested that some of these contrast dependent effects may be independent of the response saturation [56]. Nevertheless, we expect that including some type of contrast saturation in the model may improve the quantitative fit of the current model to some nCRF effects. For example, introducing contrast saturation in the surround suppression simulation (Fig. 3) may further restrict the size tuning curve peak at high contrast and lead to a closer match to the expansion ratios reported in the physiology literature. Contrast saturation could be included in this model through several mechanisms, including modifying the cost function to encourage saturating spike rates (although by itself this mechanism may not accurately capture saturating membrane potentials [89]), including LGN saturation [61], modifying the network implementation to include contrast-dependent shunting inhibition [28], or coupling the sparse coding model with a model such as the previously reported divisive normalization [65].

2.4 Materials and Methods

To implement sparse coding in a neurally plausible network architecture, we solve the dynamical system in equation (3) using a first order Euler method with an integration time step of $\Delta = 1.2$ms, 25 integration time steps per stimulus (i.e., corresponding to a stimulus presentation of approximately 1/30 second per frame of a video), a sparsity level of $\lambda = 0.5$ and a membrane time constant of $\tau = 12$ms (within the range of physiological values between 10ms and 100ms [90]). In simulations using static stimuli we measured the response after 1000 integration time steps to assure full convergence.

Stimuli such as bars and sinusoid gratings were generated as 16×16 pixel image patches, whitened (to mimic retinal processing), and overlaid on a gray background with the same mean as the gratings. Finally, for all stimuli we used a contrast (defined as the

range of the intensity values of the sinusoid grating or bar) of 0.3 unless otherwise noted.

As in physiological experiments studying nCRF effects (e.g. [3]), we first picked an arbitrary "target" neuron from the population that we would "record" from, pinpointed the center of its CRF ON-region by hand (interpreting the dictionary element as approximating the CRF), and searched for an optimal circular sinusoidal static grating patch stimulus (i.e., having the size, orientation, spatial frequency, and phase that gave rise to the maximal response of the target neuron in the model). We performed this search by a two-step exhaustive search over the parameter space using the following ranges: size of the grating was between 1 pixel and 16 pixels in diameter using 0.5 pixel increments; orientation was between 0 and 175 degrees using 5 degree increments; spatial frequency was between 0.5 to 2 radians/pixel using 0.25 radians/pixel increments; phase was between 0 to 2π using $\pi/6$ radian increments. We used this approach to map the optimal stimuli for a total of 72 simulated cells (each with CRFs well-localized within the limited visual field used in the simulation).

In most experiments we used drifting sinusoid gratings as stimuli (as described in the experimental literature for each effect). We simulated a drifting grating in discrete time by a series of static gratings at progressive phases. We fixed the temporal frequency of the grating to be about 3Hz, which is typical of the preferred frequency of cortical neurons [90]. To simulate the dynamic effect of the neural response, we simulated the dynamical system in equation (3) through the entire experiment with the driving input switched at the appropriate time to match the drift speed of the grating. We measured the response to a full cycle of the grating presentation by the mean or F1 (first harmonic) component, depending on the measure used in physiology literature for the particular effect under consideration.

In the end-stopping experiment we found an optimal static bar stimulus for the target neuron by fixing the bar width to 2 pixels, the orientation to be the same as the optimal sinusoid grating orientation, and the bar length to be the same as the optimal grating size. We then found the optimal bar location by translating the bar around a 5-pixel neighborhood

of the grating center and searching for the maximal model response for that cell. After the optimal bar stimulus location was found, we increased its length from 1 to 16 pixels and recorded the steady-state response from the model.

In the surround suppression simulation, we varied the contrast of the sinusoid grating stimuli from 0.05 to 0.5 with increments of 0.1, and we varied the size from 1 to 16 pixels in diameter with an increment of 1 pixel (other parameters were fixed). We measured the spike rate in response to the drifting grating by the F1 component. We defined the surround suppression index as $1 - a_{\text{min}}/a_{\text{peak}}$, where a_{peak} represents the peak response across all stimulus sizes at a certain contrast, and a_{min} represents the minimum response at a radius larger than the peak. Response to high contrast was measured at 0.5 and low contrast at 0.05.

In all orientation tuning studies, we stepped the orientation of the stimulus from 0 to 180 degrees in increments of 5 degrees. We measured the mean spiking response to the drifting grating. When studying the contrast invariance property, we stepped the contrast from 0.1 to 0.5 in increments of 0.1. In the population study of the tuning width, we measured tuning curve half-width at half-height by 1.17 times the standard deviation of the Gaussian fit to the orientation tuning curves. When measuring the slope of half-width vs. contrast, we normalized the contrast to 100 [8]. Five neurons in the simulated population had small unipolar CRFs and therefore showed very little orientation tunings. We could not fit Gaussians successfully to the tuning curves for these neurons, and therefore did not include their orientation tuning properties in the population study.

In the center surround orientation tuning experiment, the surround annulus grating had a thickness of 2 pixels and the center and the surround were phase-locked. When measuring the surround orientation tuning, we fixed the center orientation at the optimal orientation and measured the response to the center alone as well as the center plus the surround. We measured the response measurement for two different center radii: the optimal and the

optimal plus one pixel. In the experiment that studied the contrast's effect on the center surround orientation tuning, the center contrast took on values on a logarithmic scale (0, 0.03, 0.06, 0.12, 0.25, 0.5) and we kept the surround contrast constant at 0.5. Similar to the observation in physiology (Fig. 12a), there are many cells with weak response at low contrast in the simulation. Due to the present simulation having more cells than the study in [9], this clustering around zero made the low contrast responses difficult to read when plotted. To better visualize the suppression effect of the plaid for weakly responsive neurons, we plotted the low-contrast population responses with the maximum response normalized to 1 (effectively spreading the points out over the full range to better see their position above or below the diagonal line). High-contrast responses were similarly normalized to plot on the same scale.

CHAPTER III

MODELING INHIBITORY INTERNEURONS IN EFFICIENT SENSORY CODING MODELS[1]

There is still much unknown regarding the computational role of inhibitory cells in the sensory cortex. While modeling studies could potentially shed light on the critical role played by inhibition in cortical computation, there is a gap between the simplicity of many models of sensory coding and the biological complexity of the inhibitory subpopulation. In particular, many models do not respect that inhibition must be implemented in a separate subpopulation, with those inhibitory interneurons having a diversity of tuning properties and characteristic E/I cell ratios. In this study we demonstrate a computational framework for implementing inhibition in dynamical systems models that better respects these biophysical observations about inhibitory interneurons. The main approach leverages recent work related to decomposing matrices into low-rank and sparse components via convex optimization, and explicitly exploits the fact that models and input statistics often have low-dimensional structure that can be exploited for efficient implementations. While this approach is applicable to a wide range of sensory coding models (including a family of models based on Bayesian inference in a linear generative model), for concreteness we demonstrate the approach on a network implementing sparse coding. We show that the resulting implementation stays faithful to the original coding goals while using inhibitory interneurons that are much more biophysically plausible.

3.1 Introduction

The diverse inhibitory interneuron population in cortex has been increasingly recognized as an important component in shaping cortical activity [92]. However, it remains unclear in

[1] Key aspects of results presented here were previously published in conference abstracts [76, 91]. Full results are currently under review.

many settings how the inhibitory circuit specifically contributes to the neural code. While theoretical and simulation investigations of proposed neural coding models could be extremely valuable for providing insight into the role of inhibition, many current high-level functional and mechanistic models do not include inhibitory cell populations that approach the biophysical complexity seen in nature.

Though the main ideas likely extend to other areas, for concreteness we will focus the present discussion on the primary visual cortex (V1). In V1, visual information is encoded using a rich interconnected network of excitatory principal cells and inhibitory cells, and different coding functions appear to be implemented by distinct inhibitory populations [93, 94]. Though V1 has been extensively studied through experiment and modeling, there are often significant discrepancies between what is known about biophysical sources of inhibition and how inhibitory influences are instantiated in a model. For example, in previous high-level functional coding models (e.g. in [14, 64, 65], with the exception of [95] as discussed later), neural activity is often treated as a signed quantity without explicitly distinguishing between excitatory and inhibitory cell types. On the other hand, while state-of-the-art large scale mechanistic models (e.g. [96]) typically include a distinct inhibitory population, these types of models often use a single recurrent connectivity pattern (e.g., weights that decrease with spatial separation). This approach results in interneurons with uniform physiological properties and without the complex tuning diversity observed in inhibitory interneurons.

For theoretical and simulation studies to illuminate the role of inhibition in neural coding, it is imperative that coding models begin to incorporate experimental observations regarding the distinct properties of excitatory cells and inhibitory interneurons. Specifically, to realistically investigate the role of inhibition in neural coding, models should incorporate at least three major properties while staying faithful to the coding rule and other desirable properties (e.g., robustness):

1. Inhibitory and excitatory interactions arise from distinct cell types, and synapses

from an inhibitory cell cannot have excitatory influences on postsynaptic cells and vice versa (Dale's law [97]);

2. Excitatory neurons generally outnumber inhibitory interneurons, with E/I ratios recently estimated to be in the range $7:1$ to $6:1$ (apparently preserved across animals [98, 99]); and

3. The interneuron population has diverse tuning properties [100], including to varying degrees both orientation tuned and untuned interneurons in cat [52] and rodent V1 (reviewed in [101]).

The main contribution of this paper is to demonstrate a systematic computational method for effectively incorporating these biophysical interneuron properties into dynamical systems implementing neural coding models. In our proposed approach we exploit the fact that in many cases of interest, the total required inhibition is highly structured due to the relationship between the coding model and the statistics of the inputs being encoded. Similar to efficient coding hypotheses that postulate compact representations of sensory stimuli, the structure of the sensory statistics and the coding model can also be used to implement the required inhibition with a parsimonious computational structure. Specifically, we propose to reformulate the connectivity matrix to respect Dale's law and exploit the inhibition structure in a matrix factorization to minimize the number of inhibitory interneurons. Furthermore, we leverage recent results from the applied mathematics community on advanced matrix factorizations to develop an approach that demonstrates the observed diversity of orientation tuning properties in inhibitory interneurons.

The end result of this approach is a network implementation that is functionally equivalent to the original model, but which has an interneuron population that better respects the three major biophysical properties ignored by many current coding models. In addition to this primary goal of providing a recipe for including inhibitory interneurons into coding models, this approach also suggests possible functional interpretations of some biophysical

properties of the interneuron population. In particular, we propose that while Dale's law may reflect a physical constraint of individual cells, in contrast the E/I ratio can be viewed as an emergent characteristic of a population implementation that maximizes efficiency by minimizing the number of interneurons and thus maintenance costs. In addition, we demonstrate that the orientation tuning diversity in the inhibitory population can arise from differential connectivity patterns between the excitatory and inhibitory cells.

3.2 Results

3.2.1 Network implementation of neural coding models

In a recurrent network implementing a neural coding model, each node in the network is generally driven by both exogenous inputs (i.e., bottom-up inputs due to the stimulus or top-down feedback) and lateral connections from other cells in the same network. These lateral connections are often described in terms of a connectivity matrix G, where the element $[G]_{m,n}$ describes synaptic strength from the n^{th} neuron to the m^{th} neuron. While G can take many forms, the structure is governed by the coding model and the statistics of the stimuli being encoded.

To illustrate how G arises for a family of commonly-used coding models, we consider the Bayesian inference paradigm that has found increasing support as a framework for studying neural coding [102]. While there are many ways to develop a neural coding model based on the ideas of optimal inference, one of the most common approaches is to assume a generative model where the sensory scene is composed of a linear combination of basic features (i.e., causes) that must be inferred. Specifically, a linear generative model for vision proposes that an image patch $\mathbf{s} \in \mathbb{R}^N$ (i.e., an N-pixel image patch) can be approximately written as a linear superposition of M dictionary elements $\{\phi_i\}$ representing basic visual features (i.e., there are M principal cells):

$$\mathbf{s} = \sum_{i=1}^{M} a_i \phi_i + \mathbf{n} = \Phi \mathbf{a} + \mathbf{n}, \tag{5}$$

where the coefficients for each feature are $\{a_i\}$, $\mathbf{n}$ represents a noise source, and the $N \times M$

matrix Φ consists of one dictionary element on each column. These dictionary elements are often interpreted as the receptive fields (RFs) of a principal cell, such as spiny stellate cells or pyramidal cells.

Given the dictionary Φ and the stimulus $\mathbf{s}$, the coefficients $\mathbf{a}$ in the linear generative model (taken to be principal cell activities, such as instantaneous firing rates) can be found by maximum a posteriori (MAP) estimation. Assuming Gaussian noise and a prior distribution $P(\mathbf{a})$, the MAP estimate is found by minimizing the negative log of the posterior:

$$E(\mathbf{a}) = \frac{1}{2}\|\mathbf{s} - \Phi\mathbf{a}\|_2^2 - \lambda \log P(\mathbf{a}), \tag{6}$$

where λ is a scalar capturing the model SNR. When the prior distribution is log-concave (as are many common distributions including the exponential family [103]), the inference can be achieved by simple descent methods. The simplest dynamical system for this coding strategy would be a network implementing gradient descent with population dynamics given by

$$\tau\dot{\mathbf{a}} = \Phi^T\mathbf{s} - G\mathbf{a} + \lambda\nabla \log P(\mathbf{a}),$$

where τ is the system time constant and the $M \times M$ recurrent weight (connectivity) matrix is given by $G = \Phi^T\Phi$. G can be interpreted as a recurrent matrix because its off-diagonal terms capture the influence between cell activities. In particular when we assume that the prior is independent, i.e. $\log P(\mathbf{a}) = \sum_i \log P(a_i)$, as is common in efficient coding models, G captures *all* the recurrent influence. Note that any dynamical system involving a derivative of an energy function such as (6) will contain a recurrent matrix G of this form.

While the most obvious implementation of the network would use a single interneuron for each entry of G (connecting two cells), there are many implementations that would result in a functionally equivalent coding rule. For example, one of the approaches we will utilize is to model the connectivity between the interneurons and principal cells using a matrix factorization:

$$G = U\Sigma V^T$$

where the V^T matrix captures the synaptic connections onto a set of interneurons from the principal cells, the U matrix captures the synaptic connections from these interneurons back onto the network of principal cells, and Σ is a diagonal matrix representing the independent gains/sensitivity of each interneuron.

3.2.2 Example: Sparse Coding

As a concrete relevant example, we will demonstrate the proposed approach in the context of a dynamical system implementing a sparse coding model of V1, where a population of cells encodes a stimulus at a given time using as few active units as possible. The sparse coding model (combined with unsupervised learning using the statistics of natural images) has been shown to be sufficient to explain the emergence of V1 classical and nonclassical response properties [12, 24, 38], potentially has many benefits for sensory systems [39–41, 104], and is consistent with many recent electrophysiology experiments [29, 32, 43]. The sparse coding model has been implemented in networks that have varying degrees of biophysical plausibility (e.g., [23, 38, 47, 105, 106]), though this model has rarely been implemented with distinct inhibitory neural populations (excepting [95], discussed later).

The sparse coding model can be viewed as a special case of inference in the linear generative model described above with

$$E(\mathbf{a}) = \frac{1}{2}||\mathbf{s} - \Phi\mathbf{a}||_2^2 + \lambda||\mathbf{a}||_1, \tag{7}$$

where $||\mathbf{a}||_1 = \sum_{i=1}^{M} |a_i|$, corresponding to a Laplacian prior with zero-mean. We base our discussion on a dynamical system proposed in [23] that uses neurally plausible computational primitives to implement sparse coding. This system has strong convergence guarantees [48, 107], can implement many variations of the sparse coding hypothesis [49], and is implementable in neuromorphic architectures [50, 105, 108]. Specifically, the system dynamics for this sparse coding model are:

$$\begin{aligned} \dot{\mathbf{u}}(t) &= \frac{1}{\tau}\left[\Phi^T\mathbf{s} - \mathbf{u}(t) - (G - I)\mathbf{a}(t)\right] \\ \mathbf{a}(t) &= T_\lambda(\mathbf{u}(t)), \end{aligned} \tag{8}$$

where I is the identity matrix, $\mathbf{u}$ are internal state variables for each node (e.g., membrane potentials), $G = \Phi^T\Phi$ governs the connectivity between nodes, and $T_\lambda(\cdot)$ is the soft thresholding function. Note that despite not using steepest descent on Eq. (21), this network model still has recurrent connections described by the connectivity matrix $G = \Phi^T\Phi$. In the simulations in this study, the dictionary Φ is pre-adapted to the statistics of the natural scene with a standard unsupervised learning method, resulting in Gabor wavelet-like kernels that resemble V1 classical receptive fields [12].

This dynamical system model requires influences between cells that are described by the matrix G, but it is agnostic about the network mechanism that implements these interactions. Specifically, the model as described in [23] does not incorporate a separate population of inhibitory interneurons with any non-trivial interneuron structure, and this naïve description would only imply a point-to-point connection between all pairs of cells in the network as illustrated in Fig. 15a. This model is therefore unhelpful in its current form for understanding the coding properties of the inhibitory population.

This sparse coding network will serve as a concrete demonstration of the proposed strategy to incorporate more biophysically realistic inhibitory interneurons. The example network we use has 2048 excitatory neurons and has the same parameters as in a previous work [24] (see Materials and Methods).

3.2.3 Achieving Dale's law through factorization

As a first step towards a biologically realistic interneuron population encoding model, we show that Dale's law can be respected in the model by decomposing the recurrent connectivity matrix G into matrices representing excitatory and inhibitory interactions. Specifically, the recurrent connectivity matrix G can be decomposed into inhibitory and excitatory effects:

$$G = G_+ + G_- = G^{\text{Inhib}} + G^{\text{Excite}}, \tag{9}$$

where G_+ are the positive elements of the matrix (representing the inhibitory recurrent connections) and G_- are the negative elements (representing excitatory recurrent connections).

While G^{Excite} can be implemented by direct synapses between excitatory principal cells, the inhibitory component G^{Inhib} requires inhibitory interneurons between the relevant principal cells. To capture these disynaptic connections, we factor the inhibitory matrix into two matrices: $G^{\text{Inhib}} = UV^T$. For a simple stylized illustration, the network in Fig. 15b shows an example implementation with

$$G^{\text{Inhib}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ w_{\text{I}_1,\text{E}_3} \\ 0 \end{pmatrix}}_{U} \underbrace{\begin{pmatrix} w_{\text{E}_1,\text{I}_1} & w_{\text{E}_2,\text{I}_1} & 0 & 0 \end{pmatrix}}_{V^T}, \tag{10}$$

and

$$G^{\text{Excite}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & w_{\text{E}_4,\text{E}_3} & 0 \end{pmatrix}. \tag{11}$$

Using the approach above, we can derive a network implementation that is equivalent to the dynamical system instantiating the desired neural coding rule but that also has inhibitory cell properties that can be varied by the choice of factorization for G^{Inhib}. For a simple concrete example, we can achieve the same encoding as Eq. (8) while incorporating an inhibitory population by using the decomposition:

$$G^{\text{Inhib}} = IG_+ \tag{12}$$

where I is the identity and plays the role of U; G_+ as defined in Eq. (9) plays the role of V^T. The resulting network is shown in Fig. 16a. While this approach does utilize distinct excitatory and inhibitory sub-populations, it still requires M inhibitory neurons (i.e., one

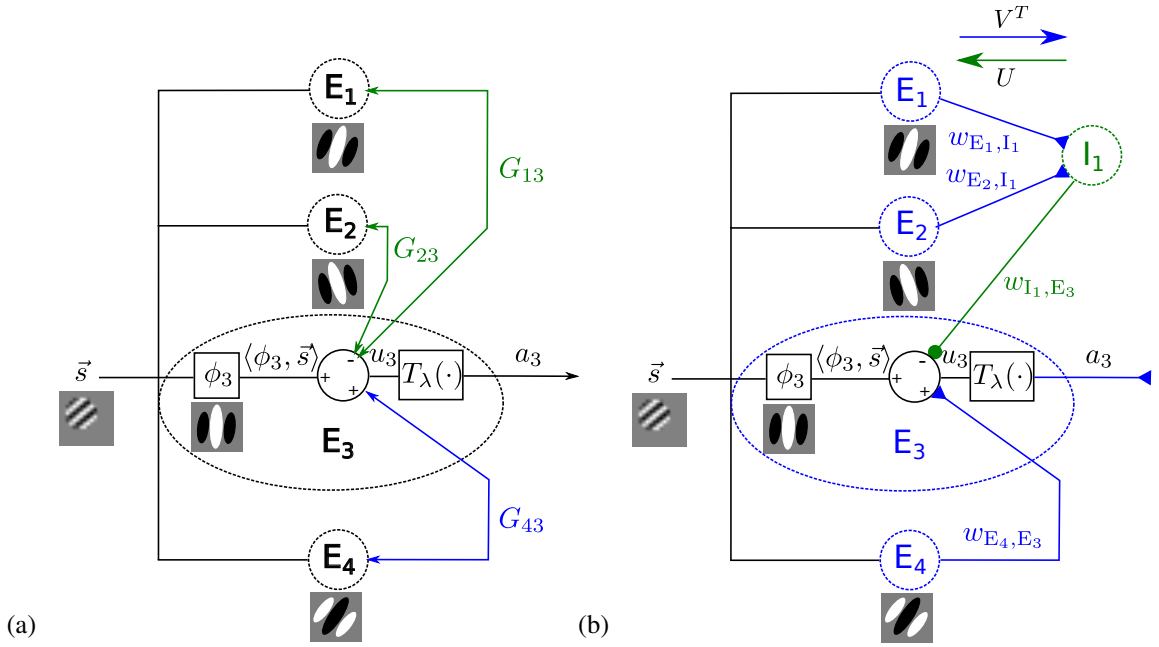


Figure 15: Achieving Dale's law. (a): An example generic neural network of visual encoding with feedforward and bi-directional recurrent connections (arrows) showing the implementation details of a single cell E_3 (other cells would be similar but are not pictured for simplicity). The sparse coding dynamics in Eq. (8) is a special case. The internal state u_3 (e.g., membrane potential) of this neuron is determined by the filtered input $\langle \phi_3, \mathbf{s} \rangle$, with the dictionary elements ϕ's depending on the natural scene statistics (e.g., [12]), the inhibitory recurrent input (green input $G_{13}a_1$ and $G_{23}a_2$ from E_1 and E_2), and the excitatory recurrent input (blue input $G_{43}a_4$ from E_4). The membrane potential is thresholded by function $T_\lambda(\cdot)$ to generate the response a_3 (e.g., the instantaneous spike rate) that drives other neurons. Note that both the excitatory and inhibitory influences are generated by the same generic cell type, violating Dale's law. (b): In this study, we incorporate distinct inhibitory interneuron populations (e.g. I_1) that are connected to the principal cells (the E-population) in specific patterns. The computational property of this type of E-I network can be shown to be equivalent to the one in (a).

for each principal cell) and all inhibitory cells in this implementation have the same orientation tuning properties as the excitatory cells (see Sect. 6.3.2). While this may introduce orientation tuning diversity due to the orientation tunings of the excitatory population, the diversity is distributed uniformly [109] instead of a bimodal dichotomy observed in the inhibitory population [13].

3.2.4 Achieving E/I ratio through low-rank decomposition

In areas such as V1, the principal excitatory cells are presumed to form the explicit representation of the stimulus that is communicated to higher cortical areas while inhibitory neurons are presumed to play a more localized computational role within a circuit. Using limited physical resources, there are many desirable properties for the stimulus representation: informational efficiency matched to scene statistics [22], stability to small stimulus changes [14], and simple downstream decoding [110]. The principal cell population in V1 appears to be substantially *overcomplete* (i.e., in both cats and primates, the estimated ratio between the output fibers and the input fibers ranges from $25:1$ to $50:1$ [17]), which is a feature adopted in some coding models because it can help achieve these desirable properties [17].

In contrast, if inhibitory neurons only need to achieve a computational goal for the circuit without requiring these same stimulus coding properties, there is no need for an overcomplete inhibitory population. In fact, the system could exploit this structure to use the fewest number of inhibitory cells possible to avoid incurring unnecessary cell maintenance costs [111]. In contrast to the direct model of Fig. 16a, this approach would require interneurons that communicate simultaneously with a *population* of excitatory neurons rather than a single excitatory neuron. As an aside, we note that the reasoning above suggests that the inhibitory population should be overcomplete in systems where these neurons *do* form the explicit stimulus representation. Indeed, this is proposed in a theory of olfactory bulb encoding where granule cell interneurons form the olfactory representation and are an overcomplete population [70].

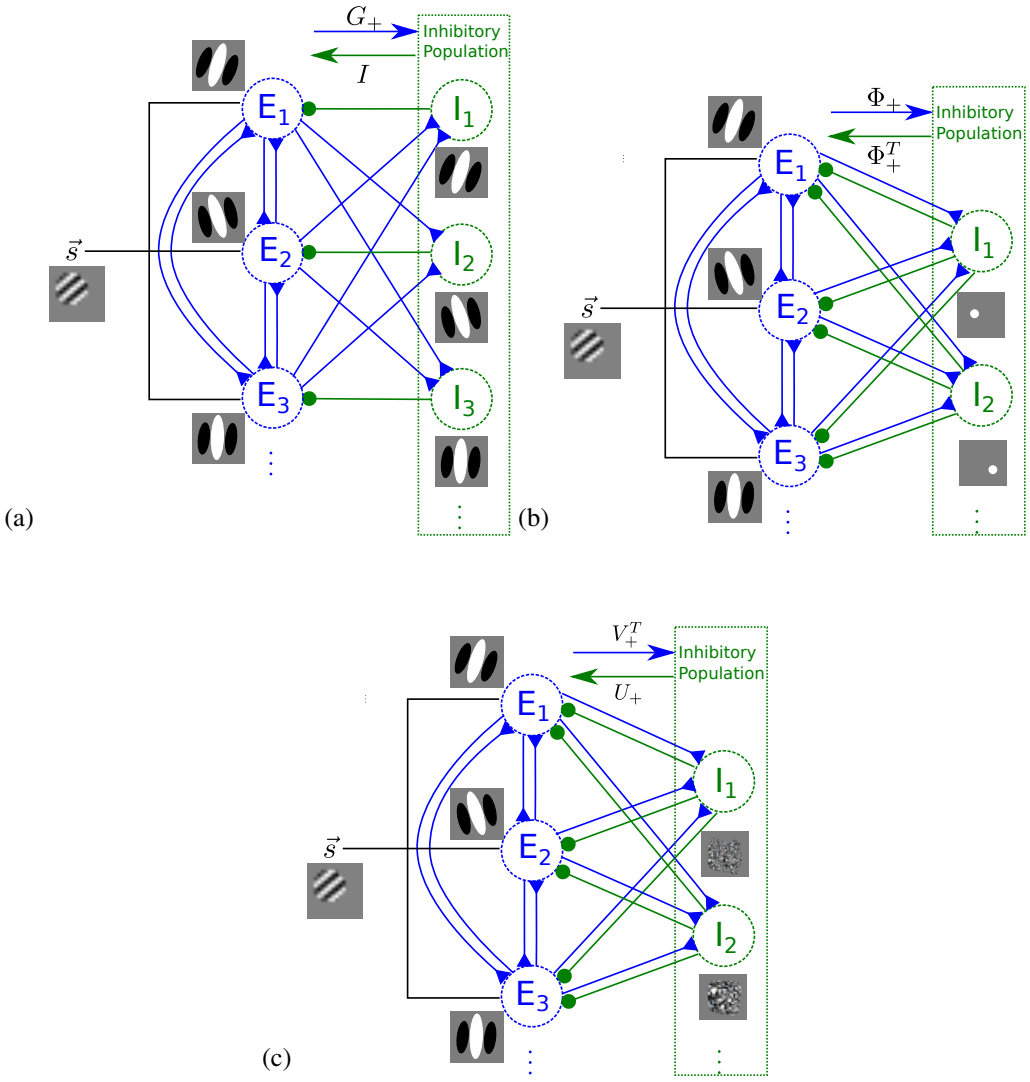


Figure 16: Achieving E/I cell ratio. (a) A subnetwork showing the connectivity and RFs in the network implementation of Eq. (12). The excitatory connection weight from E_i to the inhibitory interneurons I_j is $-\langle \phi_i, \phi_j \rangle$ (forming the $(i, j)^{\text{th}}$ entry of G_+ in Eq. (12)). The recurrent connections from the inhibitory neurons back to the excitatory ones (in green) are one-to-one (rows of the identity matrix). This implementation results in an inhibitory population with similar size and orientation tuning properties as the presynaptic excitatory cells. (b) A stylized sub-network showing the network implementing Eq. (13). The RFs (mapped out by sparse dots [12]) of the interneurons are dot-like, with extreme localization and no orientation tuning. (c) A stylized sub-network implementing Eq. (15). The interneurons receive excitatory inputs weighted by the corresponding row in V_+^T, adjust the gain by the corresponding diagonal entry in Σ, and projects back to the excitatory population with connectivity weights determined by the corresponding row in U_+. These interneurons receive dense input from many principal cells and have unstructured receptive fields, again with no discernible orientation tuning.

A natural question to ask is, what is the minimum number of inhibitory cells required to implement the influences specified by the matrix G? Said mathematically, what choice of factorization results in the fewest number of inhibitory cells, corresponding to the number of columns of U and V? In many cases of interest, the connectivity matrix G is likely to be low-rank (i.e. $M > \text{rank}(G)$), providing an opportunity to achieve an efficient implementation of the interneuron population by "compressing" the recurrent connectivity to its most essential components. There are two different causes of low-rank structure in G for the types of models considered in this study. First, an overcomplete representation of the principal cells implies directly that G is low-rank (i.e., $M > N \geq \text{rank}(\Phi) = \text{rank}(\Phi^T\Phi) = \text{rank}(G)$). Second, natural images are highly structured, meaning that image patches have fewer "degrees of freedom" than the number of photoreceptors N being used to transduce the image (i.e. $N > \text{rank}(\Phi) = \text{rank}(G)$) [112, 113]. This high level of input redundancy means that the connectivity structure implementing this coding rule also has structure that can lead to a simplified implementation. Taking both of these aspects together, models that encode stimuli with low-dimensional structure using an overcomplete code could expect to efficiently implement the encoding rule with highly-structured, low-rank connectivity matrix G.

In detail, these two sources of low-rank structure can be exploited to achieve the same coding function of Eq. (8) with fewer interneurons than a direct implementation of Eq. (12). The original description in Eq. (8) of G as a Gramian matrix gives rise to the following decomposition of the recurrent matrix:

$$G = \Phi^T\Phi = (\Phi_+ + \Phi_-)^T(\Phi_+ + \Phi_-) = \underbrace{\Phi_+^T\Phi_+ + \Phi_-^T\Phi_-}_{G^{\text{Inhib}}} + \underbrace{\Phi_+^T\Phi_- + \Phi_-^T\Phi_+}_{G^{\text{Excite}}}, \tag{13}$$

shown in Fig. 16b. Assuming first that we only take advantage of an overcomplete representation (i.e. the Φ matrix has more columns than rows because $M > N$), the resulting E/I ratio is $M : N$ and requires (potentially many) fewer inhibitory cells than excitatory cells. However, this implementation does not produce the diversity of tuning properties observed

in V1 interneurons, which can be either orientation tuned or non-orientation tuned (with no apparent structure) [13]. In fact, when using sparse dot stimuli to map out the RFs [12] of these interneurons, the resulting RFs have a dot-shaped structure (Fig. 16b) inconsistent with cortical observations (see Sect. 6.3.3 for discussion relating this RF shape to the network structure).

Further assuming that we exploit the fact that G encodes redundant structure in natural scenes, the recurrent connectivity can be represented by an even lower dimensional decomposition than Eq. (13). This can be achieved by seeking a lowest-rank (i.e., fewest number of interneurons) recurrent matrix that is also a good approximation to G (noting that up to this point we have only examined strategies that *exactly* solve the original encoding problem). Written mathematically, this approximation is:

$$L = \arg\min_{L} \operatorname{rank}(L), \quad \text{s.t. } \|G - L\|_F \leq \epsilon \tag{14}$$

where $\|\cdot\|$ is the Frobenius norm. This is equivalent to solving:

$$L = \arg\min_{L} \|G - L\|_F, \quad \text{s.t. } \operatorname{rank}(L) \leq r$$

with a suitable choice of r and ϵ. The solution to this problem can be found by the truncated singular value decomposition (SVD), known commonly as Principal Component Analysis (PCA). Note that in our case the truncated singular values are equivalent to the truncated eigenvalues because G is symmetric semi-positive definite. Specifically, we can decompose

$$\begin{aligned} G \approx L &= U\Sigma V^T \\ &= (U_+ + U_-)\Sigma(V_+^T + V_-^T) \\ &= [\underbrace{U_+\Sigma V_+^T + (-U_-)\Sigma(-V_-^T)}_{G^{\text{Inhib}}}] + [\underbrace{U_-\Sigma V_+^T + U_+\Sigma V_-^T}_{G^{\text{Excite}}}], \end{aligned} \tag{15}$$

where U and V are truncated singular vectors with orthogonal columns and implement the recurrent synaptic weights (see the Discussion section for the biological plausibility of assuming orthogonal connectivity); Σ is a positive diagonal matrix truncated from the full SVD and implements the interneuron gain (see Materials and Methods).

The resulting inhibitory population receives dense, low-rank connections from the principal cells with synaptic weights defined by V_+^T (i.e., each row representing synapses convergent onto a single interneuron) as illustrated in Fig. 16c. Note that another group of low-rank inhibitory cells with different detailed connectivity is defined by $-V_-^T$, but the qualitative characteristics of these cells are similar to those defined by V_+^T. Both groups in this population have a gain modulation defined by the diagonals of Σ, followed by projection back to the principal cells with synaptic weights defined by U_+ and $-U_-$ (i.e., each row represents synapses convergent onto a single principal cell).

In our example sparse coding network, this implementation only requires 220 interneurons to capture about 99% of the variance in G, representing a significant savings compared to 2048 and 256 interneurons required in Eq. (12) and Eq. (13) respectively. However, the resulting interneurons are again not orientation tuned, lacking the diversity observed in V1 interneurons (Fig. 16c). In Sect. 6.3.4, we show that the receptive fields of this population approximate the principal components of Φ in a generative linear model and are thus untuned.

3.2.5 Achieving tuning diversity via convex optimization

Inhibitory neurons are diverse. There are at least two populations with either tuned or untuned orientation selectivity [13]. At the same time, different inhibitory neurons connect to the excitatory population with different frequencies [114]. Could the diverse connectivity contribute to the differences in tuning? It is indeed conceivable that inhibitory neurons densely connected to the excitatory population combine inputs from different sources, and as a result have a broader selectivity. Conversely, inhibitory neurons connecting more sparsely and locally with the excitatory population might be more selective to the stimulus.

To test the hypothesis that tuning diversity could arise from differential connectivity, we decompose the recurrent connectivity matrix into two distinct matrices L and S

$$G = L + S, \tag{16}$$

where L is a dense matrix and S is a sparse matrix capturing relatively few inhibitory influences in G. To also respect the E/I cell ratio constraint, we would like L to be low-rank in particular so that a condensed representation could be achieved using SVD as demonstrated in the previous section.

Recently the applied mathematics community has developed a principled algorithmic approach known as Robust PCA (RPCA) [115–117] that effectively solves this decomposition problem. In this approach, a sparse matrix S that models "outliers" (having a disproportionate effect on the rank of G) is included so that the remainder L has a lower rank than G.

In the context of our study, an unstructured sparse connectivity matrix can result in a relatively large number of interneurons because there can be a large number of columns containing at least one non-zero value. To maintain the small number of interneurons, we also want the sparse matrix to be row or column-sparse (see for example the connectivity represented in Eq. (10)). To achieve this, we used an adaptive version of RPCA (ARPCA) [118] to decompose the recurrent connectivity matrix $G = \Phi^T\Phi$ into a low-rank matrix L and a column-sparse matrix S by solving the following convex optimization program iteratively:

$$L, S = \arg\min_{L,S} \|L\|_* + \|\Lambda S\|_1, \text{ subject to } G = L + S, \tag{17}$$

where $\|\cdot\|_*$ is the nuclear norm (i.e., the sum of absolute values of eigenvalues) to encourage L to have low rank, $\|\cdot\|_1$ is the ℓ_1-norm (i.e., the sum of absolute values of the vectorized matrix) to encourage sparsity, and Λ is a diagonal weighting matrix updated at each iteration to encourage column sparsity in S. The update rule for Λ is given by

$$\Lambda_{i,i} = \frac{\beta}{\|S^{(i)}\|_1 + \gamma}, \tag{18}$$

where $S^{(i)}$ is the i^{th} column of S, and β and γ control the speed of adaptation. At each iteration, the columns of S with smaller entries are assigned larger values of λ, thus encouraging the values in that column to become even smaller and eventually approach zero. The end

effect is that the algorithm converges to a decomposition where only a few columns in S are non-zero (see Materials and Methods for details). We note that the RPCA formulation in Eq. (17) is a natural extension to SVD in Eq. (14): instead of constraining the power in $G - L$ (via the Frobenius norm), we now model this difference using a structured matrix S.

After convergence, as before we perform a singular value decomposition (SVD) on the low-rank matrix $L = U\Sigma V^T$. To respect Dale's law we separate out the excitatory and inhibitory influence similar to Eq. (13) in each matrix:

$$\begin{aligned} G &= L + S = U\Sigma V^T + S \\ &= (U_+ + U_-)\Sigma(V_+^T + V_-^T) + (S_+ + S_-) \\ &= [\underbrace{U_+\Sigma V_+^T + (-U_-)\Sigma(-V_-^T) + S_+}_{G^{\text{Inhib}}}] + [\underbrace{U_-\Sigma V_+^T + U_+\Sigma V_-^T + S_-}_{G^{\text{Excite}}}]. \end{aligned} \tag{19}$$

With this decomposition, the recurrent matrix can be rewritten with separate excitatory and inhibitory recurrent interactions. In the sparse coding model example described earlier (Eq. (8)), the equivalent network dynamics are:

$$\dot{\mathbf{u}}(t) = \frac{1}{\tau}\left[\underbrace{\Phi^T\mathbf{s}}_{\substack{\text{feed-}\\\text{forward}}} - \underbrace{\left(\underbrace{U_+\Sigma V_+^T + (-U_-)\Sigma(-V_-^T)}_{\text{low-rank}} + \underbrace{S_+D}_{\text{sparse}}\right)}_{\text{recurrent inhibitory}}\mathbf{a}(t) + \underbrace{\left(I - G^{\text{Excite}}\right)}_{\text{recurrent excitatory}}\mathbf{a}(t) - \mathbf{u}(t)\right], \tag{20}$$

where D is a diagonal matrix with 0s and 1s on the diagonal and represents the synaptic weights on the sparsely-connected interneurons made by the principal cells (Fig. 17). With a parameter choice that strikes a balance between sparseness and low rank (see Materials and Methods), the E/I cell ratio in the model network is also close to the observed ratio. Specifically, with 2048 principal cells and 320 inhibitory interneurons (220 in the low rank population and 100 in the sparse population), the model network has an E/I cell ratio of 6.4 : 1.

Decomposing the connectivity matrix in this manner results in two distinct populations of inhibitory interneurons with a relative size controlled by the magnitude of the average

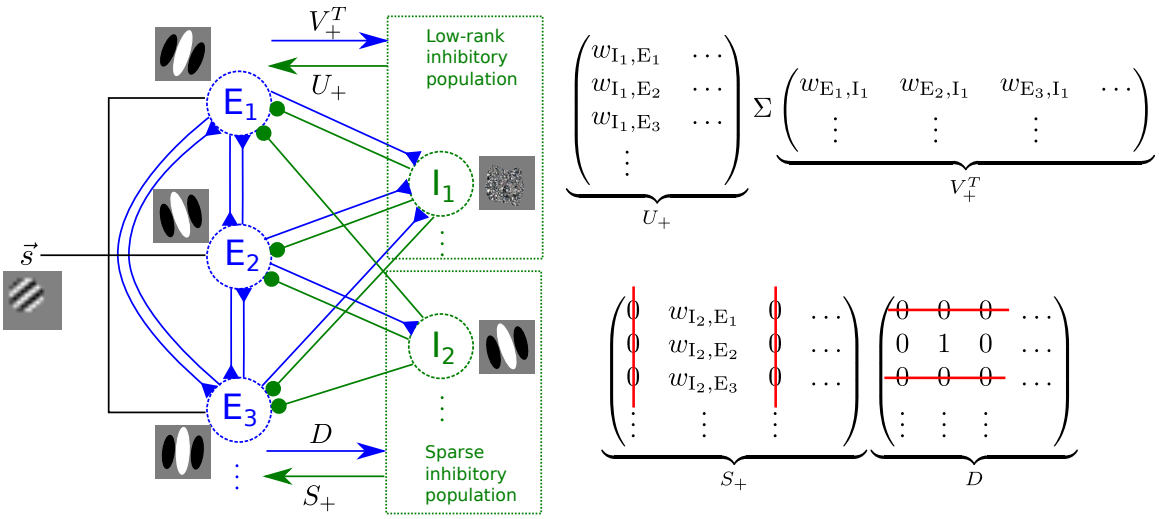


Figure 17: Low-rank plus sparse decomposition of the recurrent connectivity matrix.On the left we show a stylized network structure of the model with low-rank plus sparse decomposition of the recurrent connectivity matrix. The first inhibitory neuron I_1 belongs to the low-rank subpopulation. The second inhibitory neuron I_2 belongs to the sparse subpopulation. It receives inputs from a single excitatory neuron (E_2 in this illustration) with the connectivity matrix implemented by the diagonal matrix D, and sends projections back to the excitatory population with weights determined by a non-zero column of the connectivity matrix S_+. This inhibitory cell has the same receptive field as E_2. The matrices on the right show the decomposition of the recurrent inhibitory connections exemplified in the network on the left. The low rank and sparse inhibitory populations together implement the recurrent inhibition $-G^{\text{Inhib}}$. The excitatory recurrent influences are implemented by direct connections $I - G^{\text{Excite}}$ between the principal cells.

weights in the matrix Λ. The first subpopulation (exemplified by the inhibitory cell I_1 in Fig. 17) originates from the low-rank connectivity matrix L, and has properties described in the previous section. The second subpopulation (exemplified by I_2 in Fig. 17) originates from the sparse connectivity matrix S. This population receives one-to-one (i.e. sparse) connections with unit weights defined by the diagonal matrix D from the principal cells, and projects back to the principal cells with weights defined by S. Because S is column-sparse, the rows in D that correspond to the zero columns in S can be set to 0 without altering the recurrent influence. Said another way, we can eliminate the zero rows of the D matrix and the zero columns of S, meaning that only a few interneurons are required in this subpopulation (Fig. 17).

This model network accurately solves the sparse coding inference problem (Eq. (21)),

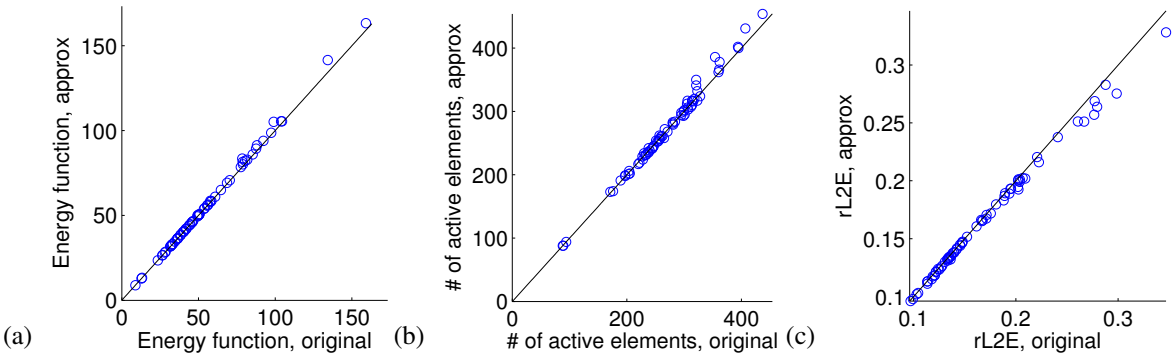


Figure 18: The network implements efficient coding. Comparison of original idealized sparse coding network model to approximation with plausible interneurons. Different markers represent results using different stimuli. (a) The energy function representing the total objective being optimized. (b) The sparsity of the response $\mathbf{a}$. (c) The relative ℓ^2 error of the image reconstruction.

despite using only the top principal components of L in the approximation to the recurrent matrix. This is shown in Fig. 27, where we compare the original network (i.e., the idealized implementation of Eq. (8) that is not biophysically plausible) with the approximation described above in the metrics of interest. Specifically, for a number of grating test patches we plot the final value of the energy function (i.e., the quantity to be minimized in Eq. (21)), along with the individual quantities relevant to the objective: the sparsity of the final answer (measured by the number of active coefficients $||\mathbf{a}||_0$) and the relatve ℓ^2 error for the input image ($||\mathbf{s} - \Phi\mathbf{a}||_2/||\mathbf{s}||_2$). As demonstrated in Fig. 27, the approximation achieves performance very similar to the original. We note specifically that in both the approximated and the original network, the activity is very sparse – only up to 5% of all 2048 neurons are active.

Interestingly, the sparse and low-rank interneuron populations in RPCA show the same kind of diverse orientation tuning as V1 inhibitory cells in vivo. The low-rank inhibitory population has RFs that are mostly untuned (Fig. 19a and Fig. 19c; orientation tuning mapped using a grating stimulus centered in the middle of the visual field), comparable to the untuned inhibitory neurons observed in cats [13] (Fig. 19b). The sparse inhibitory population has RFs that resemble the primary cell RFs in Φ and are orientation tuned (Fig. 19d

and Fig. 19f; orientation tuning mapped using a grating stimulus centered on the RF of the interneuron), comparable to the tuned inhibitory neurons observed in cats [13] (Fig. 19e). This tuning dichotomy is expected from the difference in connectivity: the orientation-tuned inhibitory RFs arise from orientation-selective inputs from single principal cells (i.e., sparse synaptic connections), whereas untuned RFs arise from dense synaptic inputs from many principal cells of different tunings.

3.3 Discussion

The main contribution of this study is a biologically plausible computational framework for including inhibitory interneurons in efficient dynamical system models of neural coding based on ideas from matrix factorization and convex optimization. From the demonstrated results, we conclude that techniques such as low-rank plus sparse decomposition can be used to find implementations of a recurrent connectivity matrix that produce equivalent population dynamics while using an inhibitory structure that matches many biophysical properties, including respecting Dale's law, known E/I cell ratios, and diversity of orientation tuning properties. In our example of a network implementing sparse coding, the resulting representation is nearly as accurate as the idealized coding model while being much more faithful to the biophysics of the inhibitory population. Because the proposed approach only depends on the structure of the recurrent matrix (which may be common among many energy based models, including many other derivatives of sparse coding [49]), we expect that the results will be applicable to many dynamical systems implementing neural coding models.

Our approach suggests that the excitatory to inhibitory cell ratio in V1 is an emergent property of interneurons implementing efficient visual coding in a resource-conserving way. Specifically, in our model a comparatively small number of interneurons efficiently route the inhibitory influence by taking advantage of the overcomplete and low-rank (redundant) structure in the recurrent connectivity pattern. We have further demonstrated that

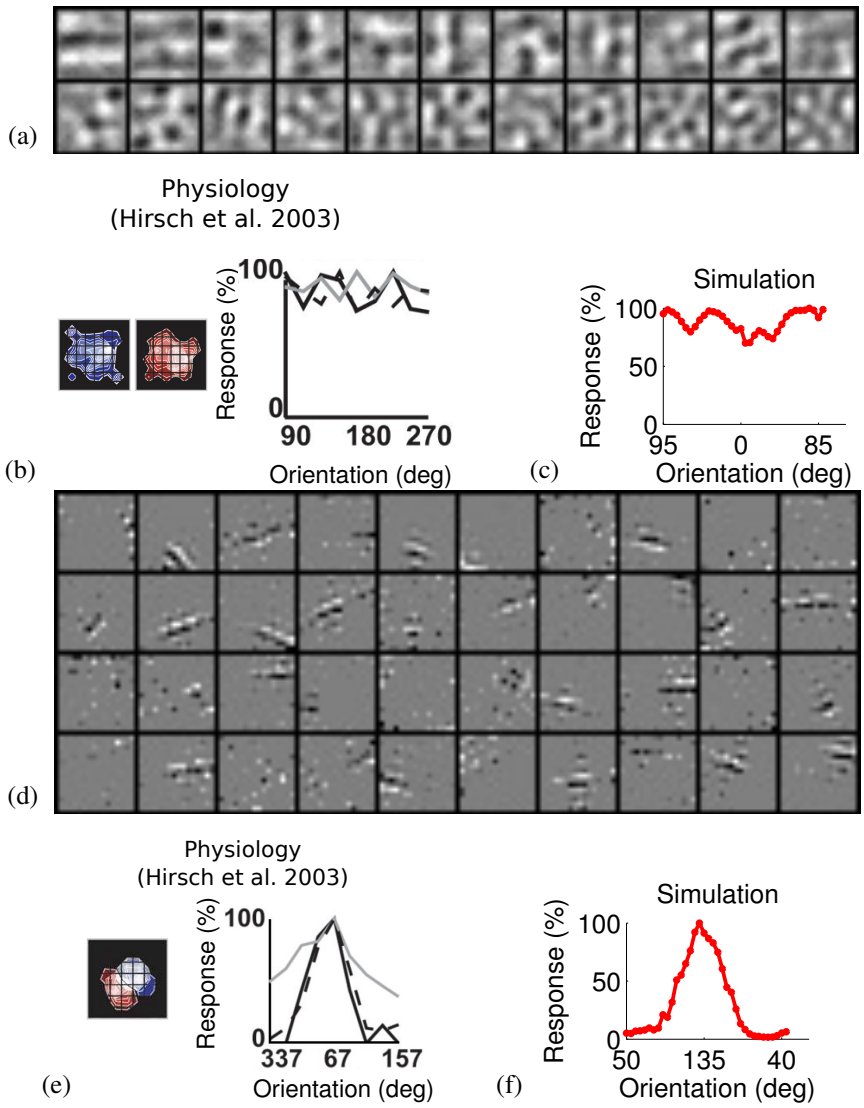


Figure 19: Achieving tuning diversity. (a) Example RFs of the low-rank subnetwork of inhibitory interneurons in the simulation. (b) An example RF and orientation tuning curve from physiological recordings (modified from Fig. 7c in [13]); (c) An example orientation tuning curve from the simulation. (d) Example RFs of the sparse subnetwork of inhibitory interneurons. (e) An example RF and orientation tuning curve from physiological recordings(modified from Fig. 4d in [13]); (f) An example orientation tuning curve from the simulation.

the tuning diversity of interneurons could arise from differential connectivity with the excitatory population – a prediction that could be tested experimentally.

3.3.1 Related studies

Recently a few studies explicitly introduced inhibitory interneuron populations into high-level functional encoding models. Lochmann et al. [68] developed a generative model that demonstrates contextual effects in sensory coding and includes a population of inhibitory neurons. These inhibitory cells contribute to efficient perceptual inference through input targeted divisive inhibition. However, this model only works with binary one dimensional inputs and the inhibitory connectivity pattern predicted by this model presently lacks anatomical support at the cortical level. Therefore, its connection with realistic visual encoding remains unclear.

In a more recent work, Boerlin et al. [119] illustrated a way to include a separate inhibitory population in an efficient coding spiking network that estimates the state of an arbitrary linear dynamical system. While providing a spiking model for the inhibitory cells, their approach did not investigate the issues of excitatory-inhibitory cell ratio and tuning diversity. It should also be noted that the Gram recurrent matrix in our model also occurs in their model (their Eq. 10). It is therefore possible that our approach could be applied in their scenario.

Another recent study [95] has developed a spiking sparse coding network based on [47] that incorporates a population of inhibitory cells with connectivity weights adapted to natural scenes. Similar to the results of our study, the work in [95] has found that a relatively small number of inhibitory cells are sufficient to provide recurrent competition required for sparse coding. In contrast, the present study formulates a framework for including biologically plausible inhibitory interneurons in a wide range of models in a way that can potentially be proven equivalent computationally to the original model objective (e.g., Eqn. (6)). Furthermore, the present work captures the observed tuning diversity of inhibitory interneurons in V1. We note that the work in [95] does use a more biophysically realistic learning

rule, whereas the present paper uses a global convex optimization approach on a fixed connectivity matrix that may have been established through a learning process.

3.3.2 Model predictions on the interneuron properties

Our model gives several experimentally verifiable predictions of interneuron properties that we detail in this section. We also note that while biologically plausible, there are limitations with the current model (see the Caveats section later).

First of all, our model predicts the existence of two distinct connectivity patterns between inhibitory interneurons and principal cells: the recurrent connections between principal cells and the low-rank interneurons are dense while the recurrent connections between the principal cells and the sparse interneurons are selective. According to these patterns, a likely biological correlate for the low-rank interneurons in mice is the fast-spiking parvalbumin-expressing (PV) interneurons, which receive dense synaptic inputs from nearby pyramidal cells of diverse selectivities [120], and project densely back to neighboring pyramidal cells [121]. Interestingly, as predicted by our model, the PV neurons indeed have broader selectivity than principal cells [122]. Similarly in cats, a subgroup of fast-spiking interneurons were found to have broader tunings than other interneurons [123]. Note that this broad selectivity means that the interneuron population derived from the low-rank component will use a dense code (i.e., most cells participating for most stimuli) even in coding rules such as the sparse coding example used in this work.

It is less clear what biological correspondence is most appropriate for the sparse interneuron population arising in the model. One candidate is the irregular firing cannabinoid receptor-expressing (CB1+) neurons, which have been shown to be more sparsely connected to the principal cells than the PV neurons [114]. However it is unclear what selectivity properties these neurons have in the visual cortex. Another candidate is the somatostatin expressing (SOM) neurons, which are orientation selective and have weaker response [122], similar to the sparse population in our model. If they indeed correspond to

the sparse population in our model, we predict that these neurons receive sparser connections from the principal cells compared to the PV neurons (this however might differ from layer to layer, as evidenced by a recent study in L2/3 [124]) .

In addition to general connectivity patterns, our model also provides predictions on the distribution of inhibitory synaptic weights in V1. As shown in Fig. 20a, we observe a near log-normal distribution of the inhibitory synaptic weights when using a dictionary adapted to the statistics of natural scenes. Compared to a standard log-normal distribution however, the model distribution has a longer tail towards the smaller values as visible from the Q-Q plot (Fig. 20b). Note that while the heavy tail is significant, in fact only a small part of the distribution deviates from log-normal (below the -2.33 quantile – corresponding to 1% of the cumulative density). Compounded with the difficulty of measuring from weak synapses, we anticipate that this tail would be hard to capture from experimental measurements. We note that there was a previous study [125] demonstrating a log-normal distribution between excitatory neurons, but we are unaware of similar findings for inhibitory cells. It should be noted that this model distribution is in agreement with the prediction of a previous model of spiking sparse coding [47]. Whether this is true in physiology requires further experimental validation.

In discussing the recurrent connections in the network of Fig. 17, we concentrate mostly on the inhibitory connections represented by the G^{Inhib} term. The excitatory influences are assumed to be implemented by direct excitatory-excitatory connections represented by the connectivity matrix $I - G^{\text{Excite}}$. The identity matrix I is assumed to be implemented by an independent mechanism that results in self-excitation. Biologically, there are at least three ways this self-excitation could be achieved: through "autapses" [126] (although most of these self-connections were observed in inhibitory cells); through excitatory interneurons that connect back to the principal cells; or through dendritic back-propagation [127].

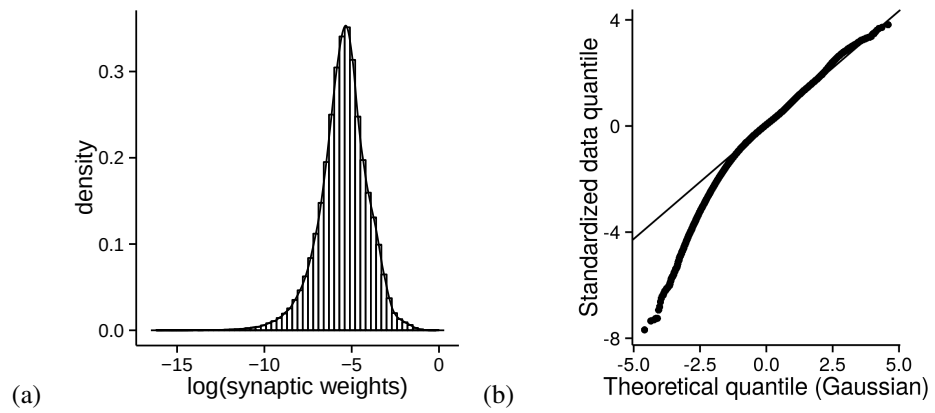


Figure 20: Distribution of synaptic weights. (a) The non-zero inhibitory synaptic weights in the RPCA model have a near log-normal distribution. (b) The Quantile-Quantile (QQ) plot of the starndardized log of the model distribution vs. a standard normal distribution. A line is drawn through the 25% and 75% quantile to illustrate the goodness of fit. The model distribution has a visible tail towards the smaller weights.

3.3.3 Caveats

We note that some of the biological features of inhibitory circuits modeled in this work are still controversial among physiology studies. For example, although Dale's law is a generally accepted operating principle, it was recently suggested that neurons can segregate neural transmitters to different synapses [128]. As another example, the diversity of tuning properties and the functional roles of inhibitory interneurons are still controversial. Most studies on this topic were conducted in rodents (the study we compared our simulation to [13] in the Results being a notable exception), with few implications for primates and leaving substantial uncertainty even in mouse neocortex [101]. For example, it is still unclear whether PV interneurons have a diversity of tuning properties [129] or are mostly broadly tuned [130]. In addition, in our simulation the recurrent inhibition sharpens the orientation tuning of the principal cells [24]; in physiology, there are conflicting accounts of whether this is the case [93, 94]. In summary, the modeling results here should be considered as a demonstration of the capability of a theoretical model to reproduce a variety of detailed biological phenomenon, not as support for any specific anatomical inhibitory circuit structures and functions.

There are several biological details of the inhibitory population that the current model does not capture. First, the non orientation-tuned inhibitory interneurons in cat primary visual cortex have complex cell characteristics such as overlapping ON/OFF receptive fields (Fig. 19b). To capture such features, a coding model involving complex cells may be necessary. Second, the current model does not attempt to capture the prevalent electrical and chemical interconnections between inhibitory interneurons in the cortex [92, 131]. These recurrent connections can potentially be incorporated by allowing off-diagonal entries in the gain matrix Σ. Third, we have treated inhibitory interneurons as continuous-time linear units with instantaneous dynamics. In reality, interneurons emit spikes and have diverse temporal dynamics involving short-term plasticity [132]. A previous work from our group [108] showed that the non-spiking sparse coding network (without a separate inhibitory population) can be equivalently implemented by a network of integrate and fire cells. While we would expect a spiking network with a similar connectivity pattern as we have demonstrated would exhibit similar kind of interneuron properties, it is unclear without further analysis whether using more biologically realistic spiking neurons would affect the overall dynamics. Finally, though it is well-known that thalamic inputs innervate inhibitory interneurons constituting feedforward inhibition [92], the model discussed in the main text does not include a detailed model of this feedforward component. However, we argue in Sect. 6.4 that the cell ratio and orientation tuning properties could be modeled in a similar manner as the recurrent network.

It is known that neural network models with different parameters may share the same input-output functionality [133]. Similarly, there are other model configurations (i.e. inhibitory connection patterns) not considered in this work that could implement the same coding functionality. For one example, in Sect. 6.5 we consider the example of global inhibition structures. In this case, while very few inhibitory cells are needed, only non orientation-tuned inhibitory cells can be modeled.

A remaining question is whether the proposed decomposition can be learned in a biologically plausible way. While it is out of the scope of the current study, we do expect the orthonormal low-rank connectivity matrices to be learnable in a biologically plausible fashion. Indeed, with Sanger's learning rule – a classical unsupervised learning method for feedforward neural networks that can be implemented locally – the network weights converge to orthonormal eigenvectors of the input [134, Chap. 8]. Note that while the orthonormality emerges automatically from the learning rule, we are not suggesting that the singular vectors are the only plausible weights in the interneuron network. For example, performing a linear transform (e.g. a rotation) in the low-rank principal subspace gives rise to an alternative decomposition that maintains the cell ratio and tuning properties we have modeled. This alternative implementation may in fact have additional computationally benefits. For example, a linear transform equalizes the gain distribution in the SVD and potentially improves the robustness of the network against noise.

3.4 Materials and Methods

3.4.1 Adaptive Robust PCA

Eq. (17) is a convex optimization problem that can be solved efficiently through numerical optimization techniques. In this study we solve this optimization problem through an adaptive version of Alternating Direction Method of Multipliers (ADMM), a robust dual ascent method [135]. Specifically, the inner loop of the algorithm finds the optimal L and S given a choice of Λ by alternating between a primal update that achieves (augmented) Lagrangian minimization and a dual update. The outer loop updates Λ according to Eq. (18). See [118] for details of the algorithm.

3.4.2 Implementation details

We start with a model network of 2048 principal neurons with receptive fields adapted to 16×16 natural image patches using sparse coding [12]. The principal cell activities are interpreted as the sparse coefficients of a dynamical system implementing sparse coding

(**a** in Eq. (8) constrained to be positive) [23] with a threshold value $\lambda = 0.1$, as was done previously in [24].

In the proposed implementation, the required number of inhibitory interneurons is governed by the rank of L and the number of non-zero columns in S. To achieve a biophysically accurate small E/I cell ratio, we would like both the rank of L and the number of non-zero columns of S to be small. However, these are two competing requirements whose tradeoff depends on the parameters in Eq. (17) and Eq. (18). Indeed, making L lower rank necessarily makes S less column-sparse. To find a compromise solution, we chose the following set of parameters: the initial diagonal of Λ is 0.038; $\alpha = 2.5$; $\beta = 0.01$. After convergence, we chose to keep 110 cells (implementing top 110 eigenvalues in L) in each of the two low-rank inhibitory populations with a total of 220 cells so that 99% of the variance in L was retained. We also used 220 interneurons in the SVD implementation to facilitate comparison between the models.

CHAPTER IV

SPARSE CODING MODELS OF POPULATION RESPONSE IN V1[1]

In the previous two chapters, we have demonstrated that sparse coding could explain many physiological properties of single cells. While providing circumstantial evidence for sparse coding as a coding strategy employed by V1, these results do not establish whether sparse coding as a population coding model could indeed account for the simultaneous activity of a population of neurons. To this end, in this chapter we analyze a multi-electrode recording data set collected in cat V1 and compare the empirical population response distribution with predictions from sparse coding. In particular, we focus on the first and second order statistics – the mean, variance, and correlation. We investigate cases where the sparse coding model presented in the previous two chapters is insufficient and propose variations of the original model that better match the observed distributions with additional coding benefits. From these models, we conclude that there are additional constraints in play in the population encoding of dynamic natural scenes in V1.

4.1 Population response to natural movies

To quantify the population response characteristics, we analyze a population activity dataset recorded in anesthetized cats viewing natural movies. The details of the experiments and the stimulus can be found in Sect. 4.8. Looking at a segment of the single-unit spike raster (Fig. 21), it is evident that the spike rates of different cells are drastically different. In the following sections, we quantify statistics such as the variance of the spike rate across cells and compare them to the sparse coding predictions.

[1]The work presented in this chapter was in collaboration with Dr. Ian Stevenson, Dr. Urs Köster, Dr. Charles Gray, and Dr. Adam Charles. Specifically, Dr. Ian Stevenson provided the spatial-temporal receptive field estimation and the movie stimulus. Dr. Urs Köster provided the single unit recording data collected in Dr. Charles Gray's lab. Dr. Adam Charles provided the base code for dictionary learning. These contributions are also noted separately in the main text.

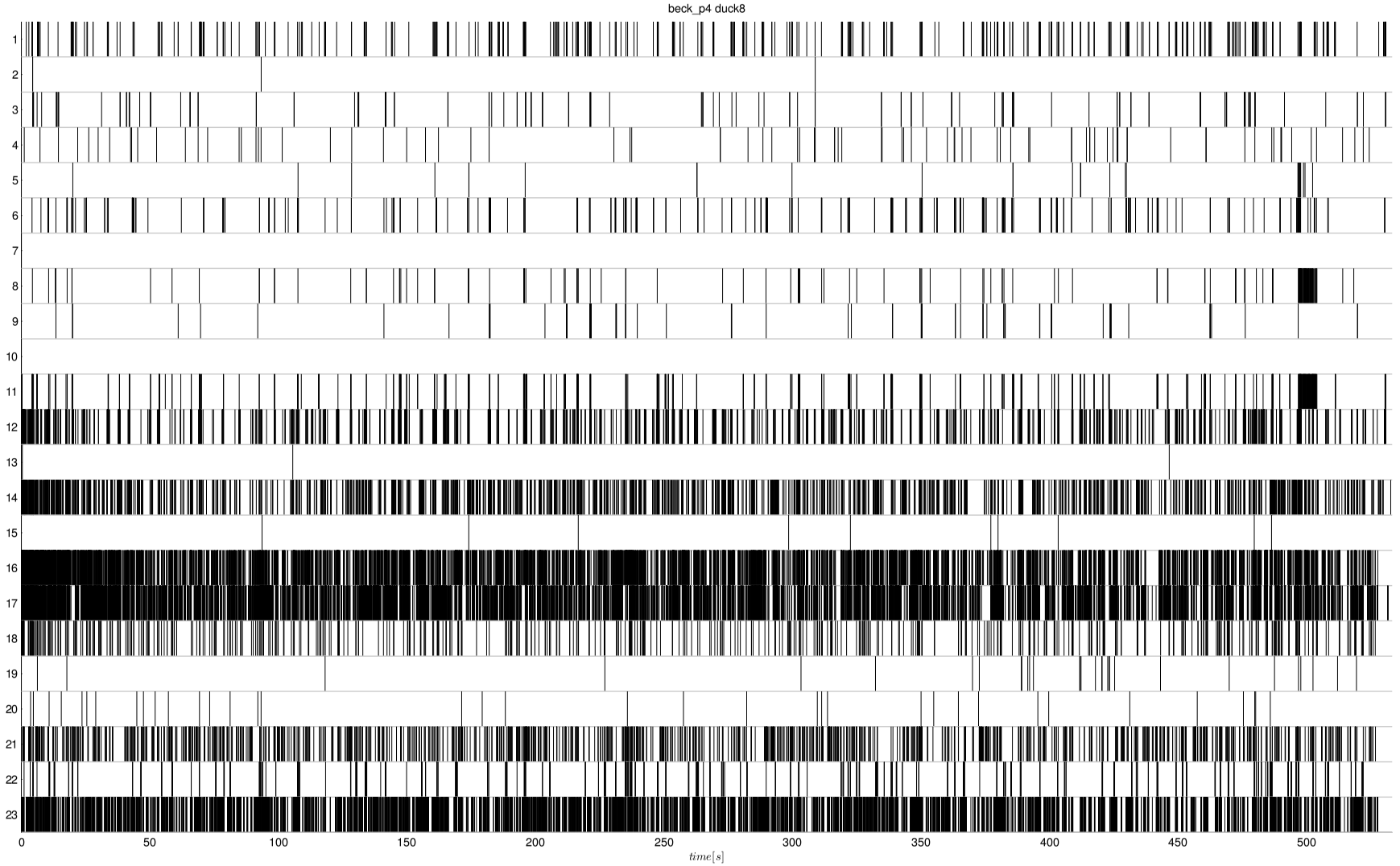


Figure 21: Single trial population spike raster responding to an approximately 10 min natural movie clip. The y-axis is the single-unit (putative cell) indices arranged roughly from the deep to the superficial layers.

Our goal is to compare the distribution of the population response predicted by sparse coding with that from the experiment. To make the measurements comparable, we make the simplifying assumption that the recorded cells in the experiment are random samples from a larger population of neurons responsive to a local visual field and that their response distribution is a fair representation of the larger population.

4.2 Training and testing sparse coding models

To compare the sparse coding prediction directly to the recorded spike activities, we need to first train the model to generate sparse code, and then further transform the model response to the same scale as the experiment.

The model training step solves the following optimization problem given the training stimulus $\mathbf{s}_{\text{train}}$:

$$\{\hat{\Phi}, \hat{\mathbf{a}}\} = \arg\min_{\{\Phi,\mathbf{a}\}} \frac{1}{2}||\mathbf{s}_{\text{train}} - \Phi\mathbf{a}||_2^2 + \lambda \sum_i |a_i|, \tag{21}$$

where $\hat{\Phi}$ is the learned dictionary and $\hat{\mathbf{a}}$ are the inferred sparse coefficients. Dictionary learning is discussed further in Sect. 4.2.1.

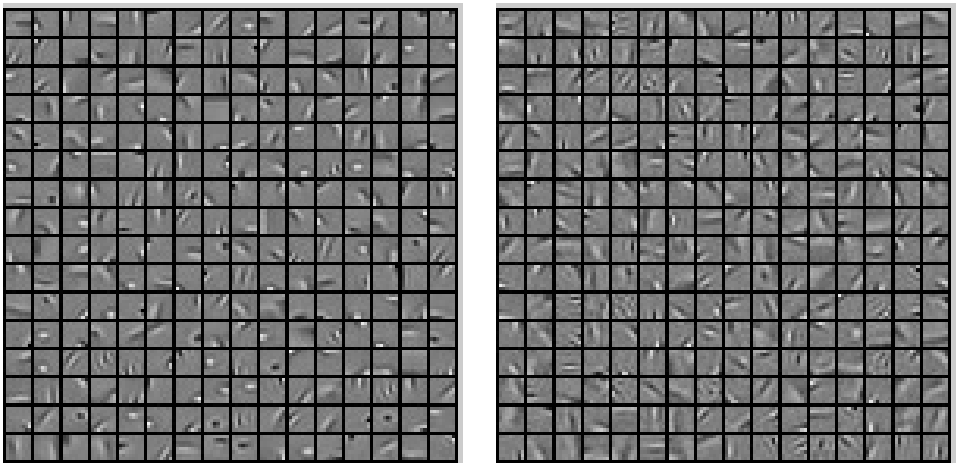
To generate sparse encoding of a new test stimulus $\mathbf{s}_{\text{test}}$, we fix Φ to the learned value and infer $\hat{\mathbf{a}}$, which amounts to solving:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \frac{1}{2}||\mathbf{s}_{\text{test}} - \Phi\mathbf{a}||_2^2 + \lambda \sum_i |a_i|. \tag{22}$$

We then transform $\hat{\mathbf{a}}$ and interpret it as the model prediction of the population response, with the details explained in Sect. 4.2.2.

4.2.1 Sparse coding learns characteristic image features from natural movie frames

The sparse coding hypothesis proposes that the neural population learns representations of the natural stimuli in an unsupervised way (Eq. (21)) so that when presented with a new stimulus with similar statistics, the population forms accurate and efficient sparse representation (Eq. (22)). To test this theory, we require that the stimulus in the test set ($\mathbf{s}_{\text{test}}$ in Eq. (22)) is similar to $\mathbf{s}_{\text{train}}$ statistically. We achieve this by using a large sample of movie



(a) Trained with natural movie frames (4x) (b) Trained with O&F natural images (4x)

Figure 22: Frames in the natural movie have "typical" natural scene local structures. (a) The dictionary trained on the natural movie clips used in the experiment is very similar to (b) the one trained using natural images in the Olshausen and Field study [14].

frames as $\mathbf{s}_{\text{train}}$ and a smaller set shot by the same camera under similar conditions as $\mathbf{s}_{\text{test}}$. See Sect. 4.8.2.1 and Sect. 4.8.2.2 for further details.

The dictionary learned using the movie frames thus prepared turns out to be very similar to one trained using static natural images in the original sparse coding study [14] (Fig. 22). This suggests that the local image statistics of the movie frames in our experiment are indeed "natural".

In fact, the two dictionaries are largely interchangeable: using one in place of another for inference does not result in a large change in the energy function in Eq. (22). To be specific, using the dictionary in Fig. 22b instead of that in Fig. 22a for inference produces a roughly 10% increase in the energy function (0.982 ± 0.019 vs. 1.117 ± 0.009; mean ± sem). This slight increase is expected since while similar, the movie frames have a slightly different statistical distribution than the Olshausen and Field images. Using the dictionary trained on one distribution to represent the other results in a slightly less efficient and accurate code.

4.2.2 Sparse coding model response is transformed to spiking events

To compare the model prediction with the experiment, we conduct simulated physiology by driving the model with the same stimulus used in the experiment (see Sect. 4.8.2.1 for further details). Yet the simulated model response $\hat{\mathbf{a}}$ in its original form is not directly comparable to the physiological measurement. First, sparse coding output is a real number that can take on both positive and negative values while the spike count in the experiment is a positive integer. Second, the scale of the sparse coefficients is different from that of the experimental measurement.

The first issue can be resolved by simulating a counting process through an inhomogeneous Poisson process generator where the rate is determined by the thresholded instantaneous sparse coefficients. Mathematically we write:

$$\begin{aligned} \hat{\mathbf{a}} &= \mathrm{T}\left(\arg\min_{\mathbf{a}} \frac{1}{2}\|\mathbf{s} - \Phi\mathbf{a}\|_2^2 + \lambda \sum_i |a_i|\right) \\ \mathbf{r} &\sim \mathrm{Poiss}(\hat{\mathbf{a}}) \end{aligned} \tag{23}$$

where T is a thresholding function that sets negative values to 0; $\mathbf{r}$ is the simulated population spike count per sample bin. In the following we use the terms *spike count* and *spike rate* interchangeably to refer to $\mathbf{r}$. We use a Poisson distribution to model the spike generation process for two reasons. First, this is a common model of the spike generation in the brain [90]. Second, we observe from the data that the overall spike count distribution indeed exhibits Poisson characteristics. As shown in Fig. 23, the variance of a given cell is roughly the same as the mean as a Poisson distribution would predict. Said another way, more active cells also have more varying spike counts across bins.

To address the second problem, we rescale the simulated response $\mathbf{r}$ with a constant factor across all neurons so that certain statistics match the experiment. For example, in the aggregate spike rate simulation, we rescaled the spike rate so that the maximum spike rates match. In the variance distribution study, the spike count variance is rescaled so that the average variance per cell is the same. This will be discussed in further detail where such

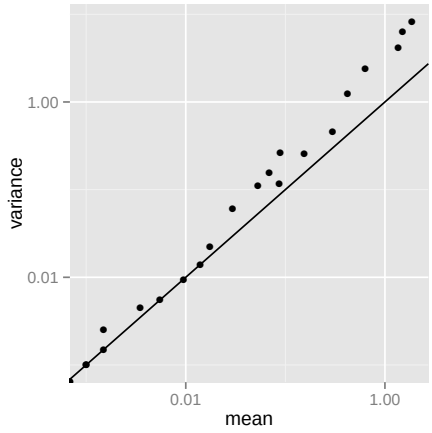


Figure 23: The spike count distribution in the experiment is approximately Poisson (mean equals variance). Each sample represents the spike count mean and variance of one neuron.

rescaling is performed.

4.2.3 Linear-nonlinear control

In the following simulations, we also include the results from the classical Linear-Nonlinear-Poisson model [136] as a control:

$$\mathbf{a} = \mathrm{T}(\Phi\mathbf{s})$$

$$\mathbf{r} \sim \mathrm{Poiss}(\mathbf{a})$$

To make the results directly comparable to sparse coding, we use the same dictionary Φ as in sparse coding. This model is similar to the sparse coding model except that it has only the feedforward filtering and does not have the nonlinear recurrent operation that sparsifies the representation.

4.3 Aggregate spike rate distribution

In this section we compare the aggregate spike rate distributions in the models and the experiment. We form the distribution by pooling spike counts from all cells and from all bins. As we see in Fig. 24, both the linear-nonlinear model and the sparse coding model (as well as a variant introduced later) capture the exponentially distributed spike rate observed in the experiment. This suggests that in both the models and the experiment there are many

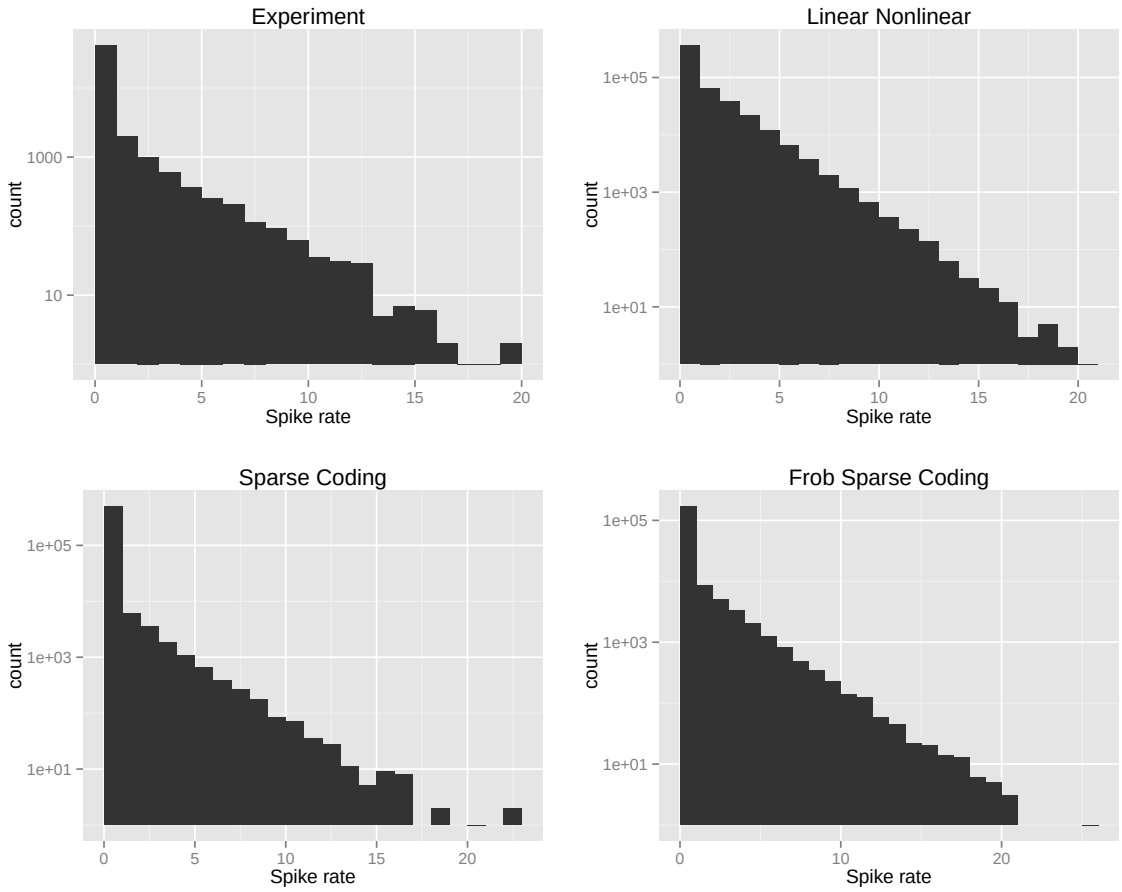


Figure 24: Comparison of spike rate distributions over all cells and all bins. Spike counts in the models are rescaled by a constant to match the approximate maximum of the experimentally observed spike rate. All the models we investigated have similar exponentially distributed spike rate as the sample distribution in the recordings.

bins containing zero spikes, i.e. the activity is sparse.

4.4 Distribution of the spike count variance

From Fig. 21, it is evident that different cells respond very differently to the natural movie stimulus – some cells do not respond at all while others fire vigorously. This diversity can be quantified by the variance of the spike count across cells. Note that due to the Poisson distributed spike count, the mean distribution is very similar (results not shown).

4.4.1 Prevalence of "silent neurons" in the recorded population

We first quantify the spike count variance distribution in the recorded population. From Fig. 25a, it is evident that cells with a close-to-zero variance are prevalent. Specifically, it is visible from the cumulative density function that around 50-75% of the population have near-zero variance/mean.

The percentage of these rarely firing "silent neurons" is in line with previous estimates. [137] showed that 40 to 80% of neurons in various cortical areas are not responsive to stimuli, at least in layer 2/3. In particular, an optical imaging study [138] revealed that only around 60% of the cells in cat visual cortex are responsive to visual stimulation. This surprising lack of activity in the cortex has also been referred to as the "dark matter" problem in neuroscience [139].

4.4.2 Equalized response in the sparse coding model

To evaluate whether the sparse coding model could account for the observed variance distribution, we conducted simulated physiology experiment on a population of sparse coding model neurons driven by the same stimulus used in the experiment (details in Sect. 4.8.2). As a control, we also carried out the same experiment on a linear-nonlinear population. We make an additional requirement that on average, the model cell variance is on a similar scale as the physiological measurement. To achieve this, we rescaled the average model variance per cell to match the physiological observation.

It is evident from Fig. 25c that while a closer match to physiology than the linear nonlinear model (Fig. 25b), the variance distribution in sparse coding is still more equalized than the physiological measurement. Indeed, the variance of many cells is low but significantly larger than zero. In addition, unlike the experiment population, every single cell in the sparse coding model has nonzero spike count mean/variance.

4.4.3 Influence of overcompleteness

A more overcomplete dictionary would result in a more selective population where different dictionary elements match different aspects of the stimulus. Given a limited set of stimuli and a very large population, it is conceivable that a small proportion of the population will never "see" the preferred stimulus and never fire.

We tested this scenario by simulating using a 16× overcomplete dictionary for sparse coding (Fig. 26). While slightly closer to the experimental observation, the distribution generated by this dictionary is nonetheless not that different from the 4× overcomplete dictionary. This suggests that something other than the level of overcompleteness is needed to account for the less equalized response in the real neural population.

4.4.4 Source of equalized variance in the sparse coding model

We contend that the equalized variance seen in the sparse coding response is a result of a hidden constraint in Eq. (21). To be specific, to ensure convergence during dictionary learning, in our simulation the sparse coding dictionary elements are normalized to 1:

$$\{\hat{\Phi}, \hat{\mathbf{a}}\} = \arg\min_{\{\Phi,\mathbf{a}\}}(\|\mathbf{s} - \Phi\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1), \text{ s.t. } \|\phi_\mathbf{i}\|_2 = 1 \tag{24}$$

This normalization has the effect of distributing the variance of the response equally among all cells. The variance-equalization was alternatively achieved more explicitly by rescaling the variance to a target constant [14].

4.5 Frobenius norm regularized sparse coding

Instead of requiring the dictionary elements to have unit norms, an alternative constraint can be applied. In particular, we can regularize the total amount of energy in the dictionary Φ through a Frobenius-norm constraint:

$$\{\hat{\Phi}, \hat{\mathbf{a}}\} = \arg\min_{\{\Phi,\mathbf{a}\}}(\|\mathbf{s} - \Phi\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 + \gamma\|\Phi\|_F^2) \tag{25}$$

The dictionary update step using this regularizer is as follows [140]:

$$\Delta\phi_\mathbf{i} \propto \langle \hat{a}_i(\mathbf{s} - \Phi\hat{\mathbf{a}}) - 2\gamma\phi_\mathbf{i}\rangle \tag{26}$$

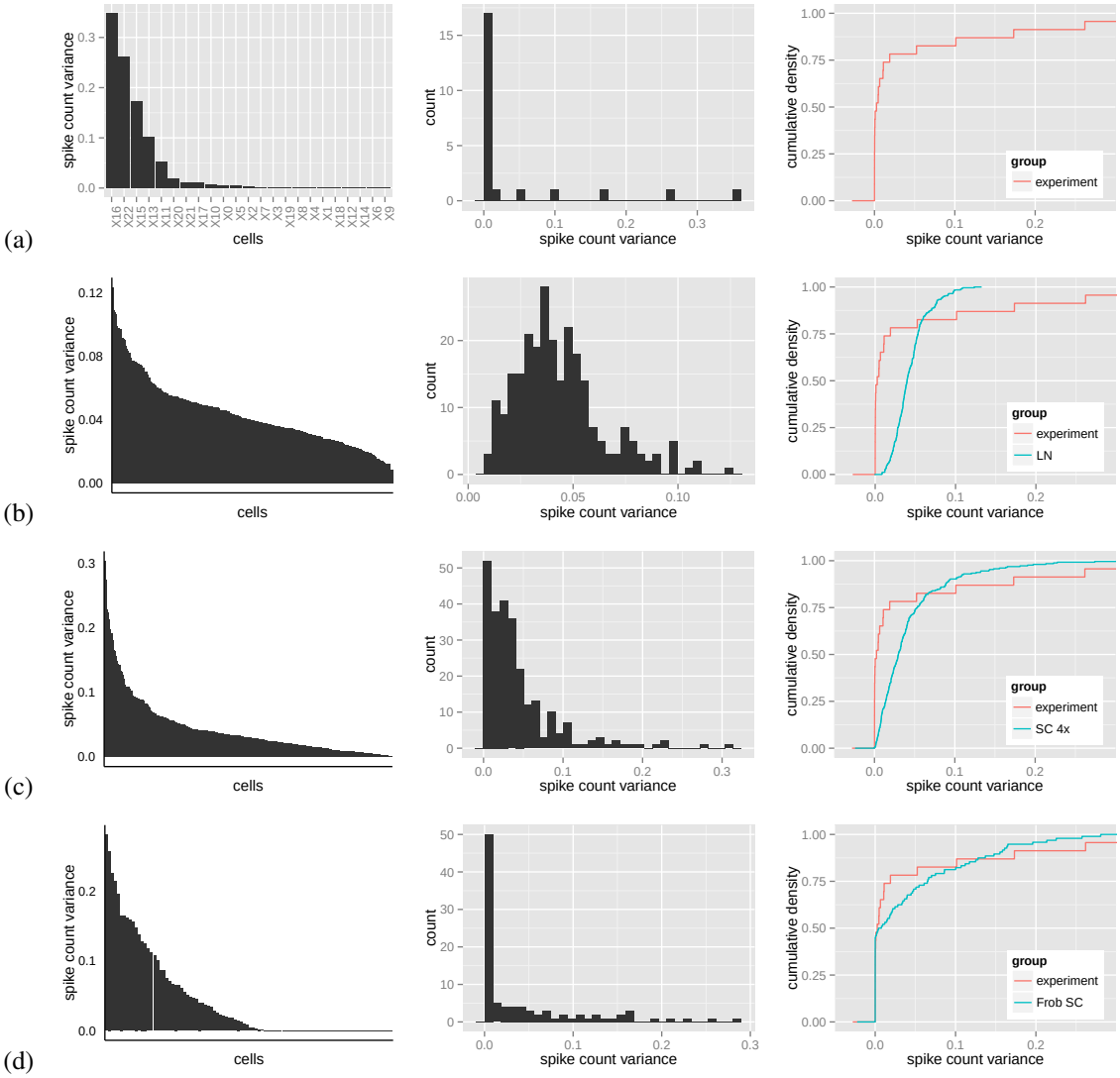


Figure 25: Comparison of the distributions of the spike count variance in experimental data and different models. The variance is scaled so that the average variance per cell is the same. Left column: spike count variance of all cells ranked from high to low; middle column: histogram of the spike count variance; right column: empirical cumulative distribution function of the spike count variance. (a) Experiment; (b) Linear nonlinear model; (c) 4 times overcomplete sparse coding model; (d) 1.5 times overcomplete Frobenius norm regularized sparse coding model.

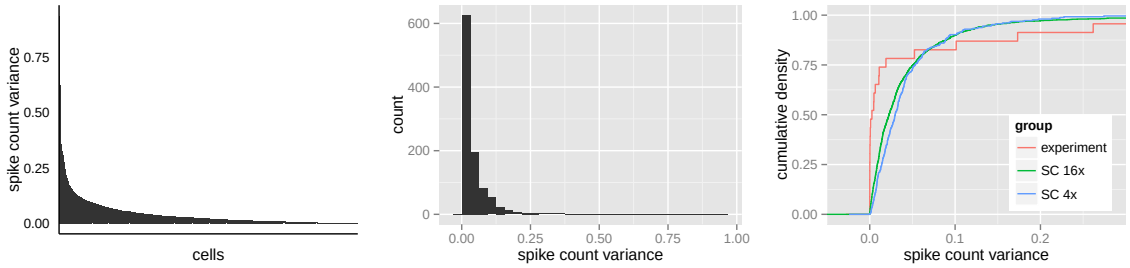


Figure 26: Effect of overcompleteness on the spike count variance distribution in sparse coding models. Increasing the overcompleteness by 4 times does not change the empirical cumulative distribution function much. Note that the scale of the left and the middle figures is different from Fig. 25.

At the steady state, this update equation suggests that a dictionary element with a smaller amplitude also has a lower activity level on average and vice versa. An expected consequence is that the dictionary elements used less often will get smaller and eventually approach zero. This is indeed what we observe in the dictionary learned under the Frobenius-norm regularization (compare Fig. 28b with Fig. 22a). Due to this diversity in the learned dictionary elements, we expect the variance of the response in the Frobenius-norm population to be more heterogeneous.

4.5.1 Frobenius-norm regularized sparse coding better fits the sample variance distribution

When inferring the model response by solving Eq. (22), switching from a dictionary learned from the original sparse coding to one with Frobenius-norm regularization results in a more heterogeneous spike count variance distribution (Fig. 25d). As shown in the figure, this distribution in fact matches the experiment better. More quantitatively, using the Kolmogorov-Smirnov distance measure, at $p = 0.01$ level the variance distribution in a sparse coding model is significantly different from that in the experiment while the distribution in sparse coding with Frobenius-norm regularization is not (Tab.1).

Table 1: Comparison of the K-S distances. At $p = 0.01$ level, we cannot reject the hypothesis that the sample distribution of the variance comes from a Frobenius-norm model.

	K-S distance	p-value
Linear-nonlinear	0.727	3.997e-10*
Sparse coding (4x)	0.543	7.948e-06*
Sparse coding (16x)	0.4598	1.5e-04*
Frob-norm (1.5x)	0.361	0.01589

Table 2: The total synaptic weights in the Frobenius norm regularized sparse coding model is smaller than that in the original sparse coding model.

	$\|\Phi\|_1 + \|\Phi^T\Phi\|_1$
Sparse coding (4x)	8.5e3
Frob-norm (4x)	2.8e3

4.5.2 Frobenius norm regularized sparse coding reduces total synaptic weights

In the network implementation of sparse coding [23], the synaptic weights scale with the dictionary Φ:

$$\dot{u}_i(t) = \frac{1}{\tau}\left[\underbrace{\langle \phi_i, \mathbf{s}\rangle}_{\substack{\text{feed-}\\ \text{forward}}} - u_i(t) - \sum_{j\neq i} \underbrace{\langle \phi_i, \phi_j\rangle}_{\text{recurrent}} a_i(t)\right] \tag{27}$$

$$a_i(t) = T_\lambda(u_i(t))$$

More precisely, the total absolute value of the synaptic weights in this network is $\|\Phi\|_1 + \|\Phi^T\Phi\|_1$. Because the Frobenius-norm regularizer encourages smaller Φ, it is expected that the total synaptic weights in the Frobenius-norm model is smaller than the original sparse coding model. This is indeed the case (Tab.2).

A smaller total synaptic weights is desirable biologically because it reduces the total synaptic volume – a costly resource [141]. Interpreted this way, the Frobenius-norm regularizer naturally incorporates an additional resource constraint into the sparse coding framework and gives rise to a more efficient code.

4.5.3 Frobenius-norm regularized sparse coding learns a "better" dictionary

We can assess how good a representation a learned dictionary is by evaluating the objective (energy) function in Eq. (22) during inference. A better representation affords the model

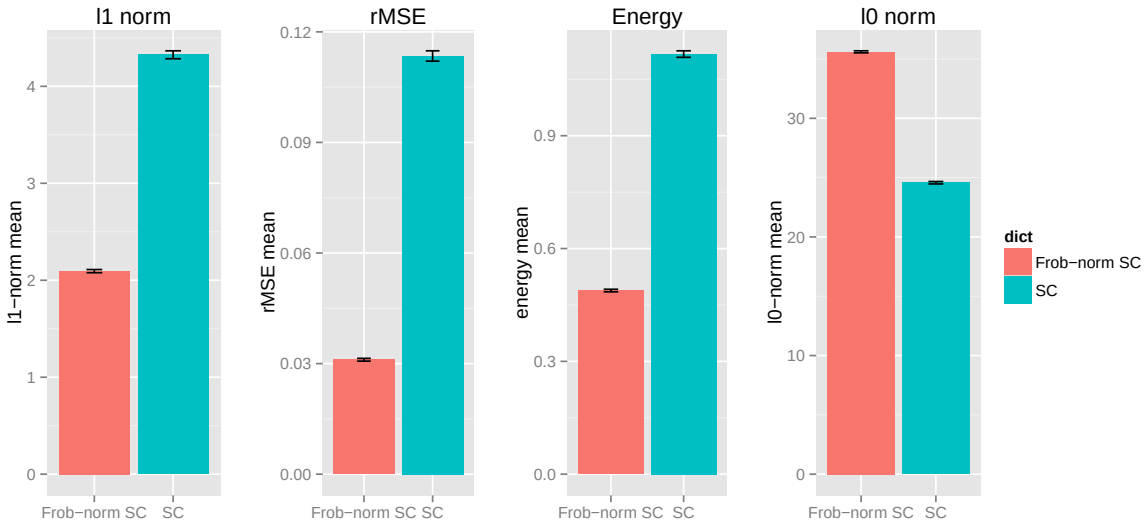


Figure 27: Comparison of MSE and sparsity measures. The dictionary learned with Frobenius-norm regularization leads to inference that achieves lower rMSE and higher sparsity and a consequent lower energy compared to the original sparse coding. The sparsity measured through ℓ^0-"norm" is slightly higher although this measure depends on what threshold we choose to measure ℓ^0-"norm".

to find more accurate and/or sparser solutions to Eq. (22). Using the objective/energy function as a measure, in Fig. 27 we show that the Frobenius-norm regularized sparse coding is a "better" model than the original sparse coding dictionary. In fact on average, the Frobenius-norm regularized version learns a dictionary that achieves both a lower rMSE and a smaller ℓ^1-norm on the test set. We note that the Frobenius-norm regularization does increase the ℓ^0-"norm" – the number of non-zero elements – on average, here measured as the number of neurons with response greater than $1e^{-4}$ given a stimulus. With a higher threshold however the Frobenius-norm regularized version has a lower ℓ^0-"norm". For example, with a threshold of $1e^{-1}$, Frobenius-norm version has a mean ℓ^0-"norm" of 6.21 while the original sparse coding has a mean norm of 13.4. This is because Frobenius-norm version has fewer large coefficients on average compared to the original sparse coding.

4.5.4 Frobenius-norm dictionary size adapts to the training set complexity

Regularizing with a Frobenius-norm constraint has the added benefit that the effective dictionary size does not have to be pre-specified as in the original sparse coding. The learning process in Eq. (26) automatically scales the unused dictionary elements to zero, dropping them off from the representation.

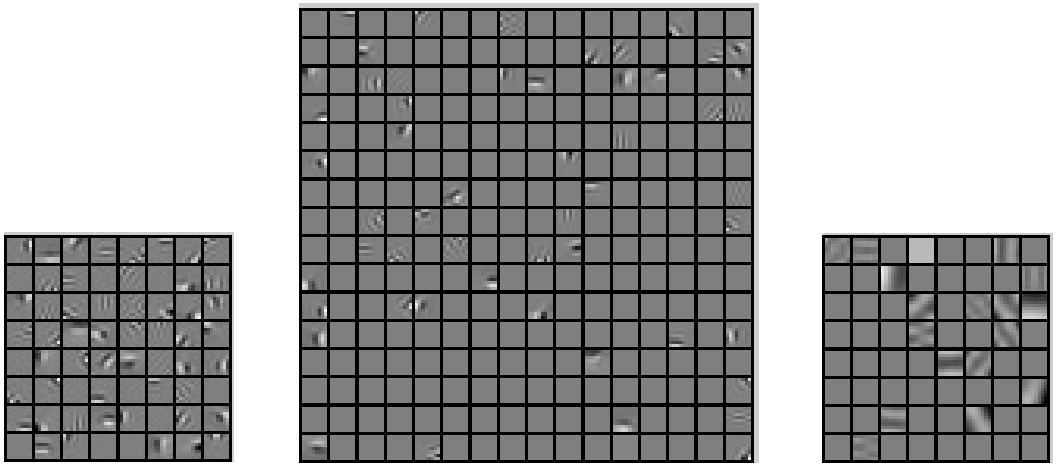
To illustrate this point, we compare the effective dictionary sizes (number of non-zero dictionary elements) at different levels of overcompleteness (Fig. 28a vs. Fig. 28b). It is evident that when trained on the same data, the dimensionality of the learned representation stays essentially the same independent of the prescribed dictionary size.

We further demonstrate that the effective size of the dictionary learned with the Frobenius-norm regularizer reflects the complexity of the training set. Fig. 28c shows a dictionary learned from a set of stimuli with only the low-frequency content (prepared by "zooming in" on the local details of the movie frames; see Sect. 4.8.2.2). Compared with a dictionary trained with more complex stimuli (Fig. 28a), the low-frequency dictionary has a much smaller effective size.

4.5.5 Biological relevance

The Frobenius-norm regularized sparse coding suggests that silent neurons observed in the experiment are vestiges of a learning/evolutionary process and do not contribute to coding. Maintaining these neurons presumably incurs a cost, so why are they not optimized away? A speculative explanation is that these neurons form a "reserve" that makes the population more robust. Our unsupervised learning model suggests that these silent neurons will become active when some of the original active neurons die, or when changes in the environment require new stimulus statistics to be represented. This "revival" of silent neurons likely involves dendritic plateau potentials [142] instead of back-propagation of action potentials as the post-synaptic cell does not have somatic spikes.

There are many other factors that may account for the existence of silent neurons. First of all, it may well be that silent neurons are in fact not silent when presented with a suitable



(a) Size: 53 (b) Size: 54 (c) Size: 18

Figure 28: Effective dictionary size of the Frobenius-norm regularized dictionary adapts to the training set. (a) Frobenius-norm regularized dictionary with 64 elements and 53 non-zero elements. (b) Increasing the dictionary size to 256 does not significantly influence the number of non-zero dictionary elements. (c) Trained on less complex low spatial frequency natural image patches, the effective size of the Frobenius-norm regularized dictionary is lower.

stimulus. In other words, many neurons in the visual cortex may be so selective that only a very specific stimulus can induce a response. Our result suggests that this selectivity is unlikely to arise from sparse coding alone even with a highly overcomplete dictionary (Fig. 26). Another potential source of silent neurons is anesthesia. Anesthesia is known to change the state of arousal and the excitability of neurons [143] and may "turn off" a subpopulation of neurons that would be responsive otherwise. However it is unclear how strong this effect is since silent neurons are observed in awake animals as well [137] and the overall distributions of firing rate are similar under awake and anesthetized conditions [21].

Note that the computational benefits of Frobenius-norm regularized sparse coding demonstrated in Sect. 4.5.2 to Sect. 4.5.4 are independent of the silent neurons. In particular, silent neurons do not contribute to either the accuracy or the sparsity of the representation. Furthermore, when counting the dictionary size that adapts to the stimulus complexity, the silent neurons are not considered as part of the population.

4.5.6 Predictions

Our model offers several predictions that can be validated by experiments. First, we predict that the synapses onto the silent neurons are weak. A consequence of this is that these neurons would be extremely hard to detect with electrophysiology even with direct presynaptic stimulation. It is perhaps necessary to rely on optical imaging methods (e.g. [138]) to detect the silent neurons, and our prediction can then be validated by directly measuring synaptic strengths in these detected neuron.

A second prediction of the model is that animals reared in a visually deprived environment where stimuli have simple structures will develop more silent neurons. Conversely, allowing more complex stimuli in the animal's visual experience (especially during the critical period) may increase the size of the active population (Fig. 28).

4.5.7 Future works

Biological mechanisms other than the differences in synaptic weights may contribute to the observed response variability. For example, a sub-population of inhibitory cells are fast-spiking [120] and may partly account for the high-variance population in the recording. This factor could be introduced easily into our model without compromising the computational properties using the techniques introduced in Chap. 3.

A remaining question is how the learning rule in Eq. (26) could be implemented in a biologically plausible way. A recent work in online sparse coding dictionary learning with biologically plausible networks may offer a solution [144].

4.5.8 Conclusion

It is evident from the recording that some neurons in V1 are highly active while others are completely silent. How do we explain the observed diversity from a coding perspective? The Frobenius-norm regularized sparse coding model provides a tentative answer. Specifically, this model suggests that the highly varying activity level is a consequence of an

adaptive population minimizing the representational cost (population firing rate and synaptic weights) while maintaining the accuracy. The synaptic weights constraint is crucial in explaining the diversity in the observed variance: without this constraint, the original sparse coding model does not predict the observed activity distribution well. Incidentally, introducing this constraint results in a more sparse and accurate representation and a more flexible population that self-adapts to the complexity of the training stimuli.

4.6 Correlation structure

We have investigated thus far the statistical structure of the spike rate distribution in individual neurons as if they were independent. The neural population, however, have structured connectivity patterns and exhibit characteristic correlation structures. It is currently unclear how well a theoretical model like sparse coding could account for the empirical correlation structure and this will be the focus of this section.

4.6.1 Correlations in the physiology experiment

From the pairwise Pearson's correlation matrix (see Sect. 4.8.2.4 for details) between the trial-averaged response (Fig. 29a), it is visible that there are clusters of neighboring cells showing strong correlation, especially in the superficial layers. This is also visible in the single trial response (Fig. 29b), albeit less strongly.

Could receptive fields explain the observed response correlation? Fig. 29c shows an interesting relation between the response correlation and the receptive field similarity (estimated by spike-triggered average; see Sect. 4.8.1.3): while cells with similar receptive fields also have highly correlated response, the response correlation in cells with dissimilar receptive fields are close to 0. Note that here for consistency we show the correlation distribution for the same population we have studied so far in this chapter. This distribution is in fact quite invariant across animals and experiments (Sect. 6.7).

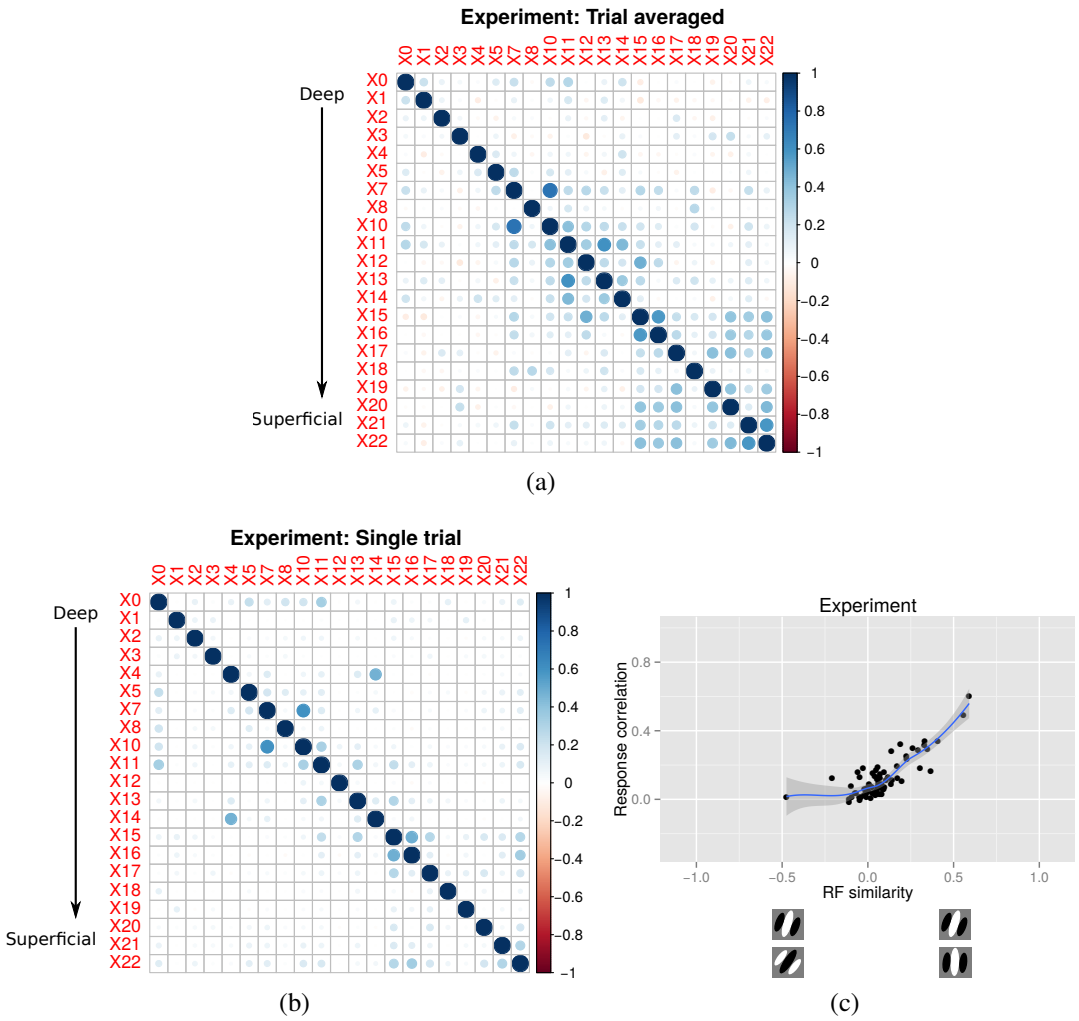


Figure 29: Response and receptive field correlations between 21 cells in the experiment data (two silent neurons were excluded) (a) Correlation matrix of trial averaged response (PSTH) with cells arranged roughly from deep to superficial layers (b) Correlation matrix of single trial response (c) Relation between the response correlation and the receptive field similarity in a single trial. The solid line is a local polynomial regression fit to the scatter.

4.6.2 Correlations in the linear-nonlinear and sparse coding models

With simulated physiology (Sec. 4.8.2) we find that neither the sparse coding model (Fig. 31) nor the linear-nonlinear control model (Fig. 30) captures the correlation pattern observed experimentally. In particular, the linear-nonlinear model predicts a purely linear relation between the response correlation and the receptive field similarity, while the true response correlation saturates near zero. On the other hand, sparse coding predicts a close-to-zero response correlation no matter how similar the receptive fields are[2].

In addition, the connectivity patterns predicted by these two models are not in line with the emerging view of the layer 2/3 circuit in the visual cortex [145–147][3]. To be more specific, recent studies suggests that in layer 2/3, excitatory cells with similar receptive fields preferentially connect with each other. Contrary to this highly structured recurrent circuit, the linear-nonlinear model does not predict clusters of neurons with similar selectivity, while sparse coding predicts that excitatory cells that preferentially connect with one another have *dissimilar* receptive fields.

4.7 Group sparsity model

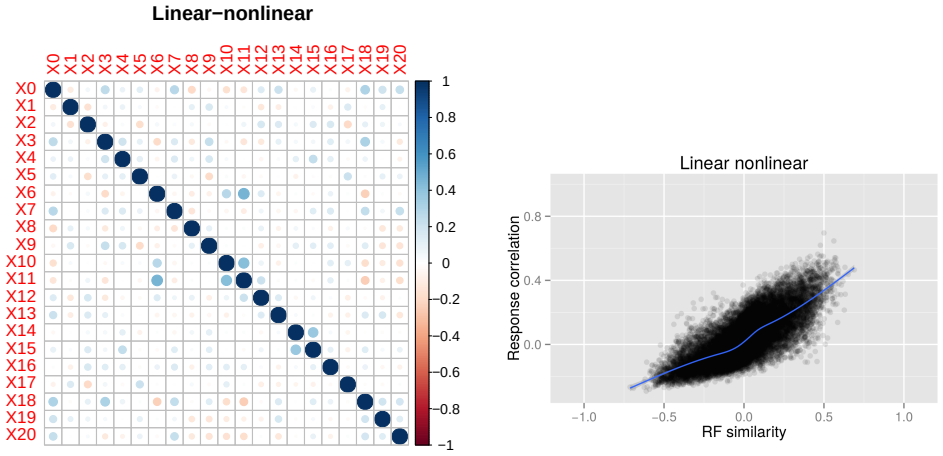
A potential way to encourage clusters of correlated neurons to emerge in a sparse coding model is to use a *group sparsity* constraint in place of an ℓ^1-norm constraint:

$$\{\hat{\Phi}, \hat{\mathbf{a}}\} = \arg\min_{\{\Phi,\mathbf{a}\}}(\|\mathbf{s} - \Phi\mathbf{a}\|_2^2 + \lambda \sum_j \|\mathbf{a^j}\|_2), \tag{28}$$

where $\mathbf{a^j} = (a_1^j, a_2^j, ..., a_{n_j}^j)$ represents the activity in the jth group. This alternative mixed ℓ^1/ℓ^2 norm [149, Chap. 2] encourages a small number of groups to be active for a given stimulus and places no constraints on the sparsity within these groups. With this alternative formulation, a grouped representation can be learned with each group representing a "subspace" of the training set. Resulting dictionary elements within the same group appear to have similar orientation selectivity and different phase selectivity (Fig. 32).

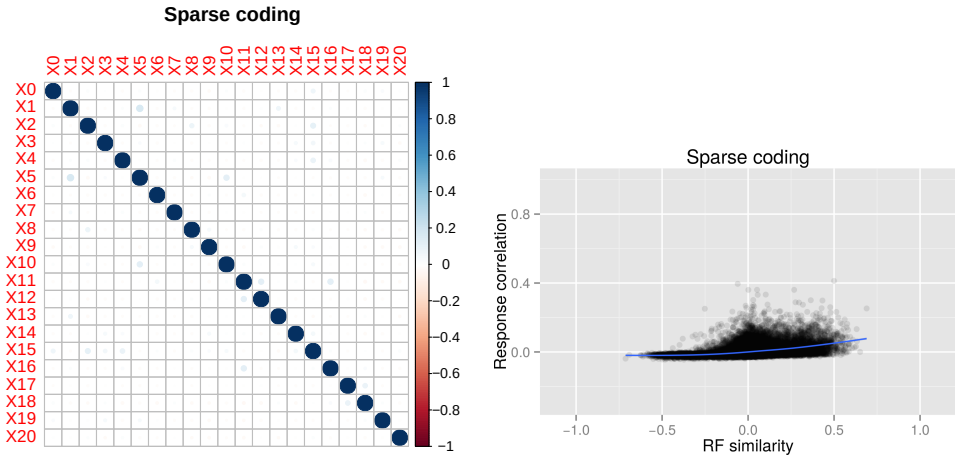
[2]Frobenius-norm regularized sparse coding predicts a similar relation (not shown).

[3]These studies were conducted in rodents. It is currently unclear how the wiring diagram in cats differ [148].



(a) (b)

Figure 30: Response and receptive field correlations in the linear nonlinear model. (a) Response correlation matrix between the first 21 model neurons in a simulated single trial. (b) Relation between the response correlation and the receptive field similarity in all 256 cells. The solid line is a generalized additive model fit to the data.



(a) (b)

Figure 31: Response and receptive field correlations in the sparse coding model. (a) Response correlation matrix in a simulated single trial. (b) Relation between the response correlation and the receptive field similarity. The solid line is a generalized additive model fit to the data.

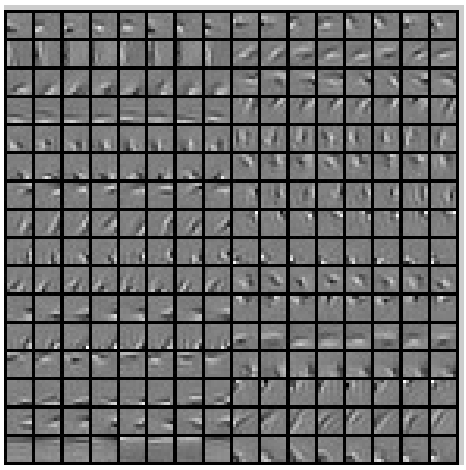


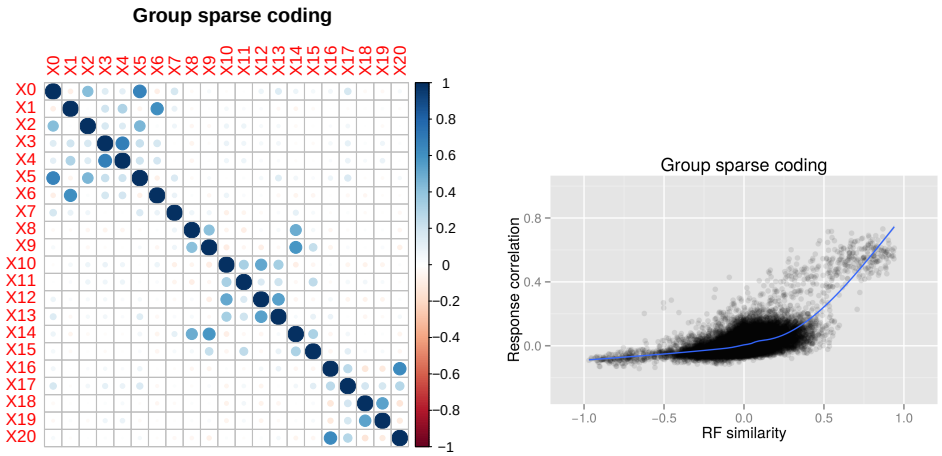
Figure 32: 4x overcomplete dictionary of size 256 learned with a group sparse prior of group size 8

4.7.1 Group sparse coding better captures empirical correlation

Comparing the group sparse coding prediction of the correlation pattern (Fig. 33a) with the experiment (Fig. 29a and Fig. 29b), we see that the model prediction indeed resembles the clustered correlation pattern observed in the experiment. In addition, the group sparse coding model matches the experimental observation that the response correlation increases with the receptive field similarity nearly linearly when the receptive fields are positively correlated, and that the response correlation stays near zero when the receptive fields are dissimilar (Fig. 33b).

We can interpret these results by comparing to sparse coding and linear-nonlinear models. Unlike in sparse coding, the response correlation between cells with similar receptive fields is much larger in the group sparse coding model. This is due to the fact that cells within a group represent the same subspace and both their response and their receptive fields tend to be correlated.

Compared to the linear-nonlinear model, in group sparse coding the response correlation between neurons with anti-correlated (opposite phase) receptive fields is much closer to zero. One explanation is that in the group sparse coding model, cells with dissimilar receptive fields tend to be in the same group. For these cell pairs, the positive correlation



(a) (b)

Figure 33: Response and receptive field correlations in the group sparse coding model. (a) Response correlation matrix of a subset of cells in a simulated single trial. (b) Relation between the response correlation and the receptive field similarity. The solid line is a generalized additive model fit to the data.

induced by the group structure partially cancels out the negative correlation, leading to an overall correlation close to zero.

To quantify the difference between the the model distribution and the experimental distribution, we measure the statistical distance defined as the integrated squared difference between the kernel-smoothed densities:

$$T = \int [f_{\text{exp}}(\mathbf{x}) - f_{\text{model}}(\mathbf{x})]^2 d\mathbf{x}], \tag{29}$$

where $f(\mathbf{x})$ is a Gaussian kernel-smoothed density function estimated from the data (see Sect. 4.8.2.4 for details). As seen in Tab.3, the group sparse coding model with a group size of 8 produces a distribution that that best matches the experimental data while the group sparsity model with a smaller group size as well as the linear-nonlinear and the original sparse coding model generate less accurate predictions. While the distribution generated from a group-of-8 sparse coding model is the closest to the sample distribution, a statistical test revealed that the model distribution is significantly different from the sample distribution (p-value exceedingly small). More refined models are needed to match the sample distribution even better (Sect. 4.7.4).

Table 3: Comparison of the statistical distances from the model distribution to the experiment distribution. Note that to better estimate the distribution of the experiment data, here we incorporate auxiliary experimental data from other sources (see Sect. 6.7).

Model	Integrated dist. to exp. distr. (T)
Sparse coding	61.42
Linear-nonlinear	29.10
Group-of-4 SC	27.88
Group-of-8 SC	24.35

4.7.2 Group sparse coding as a model of complex cells

A classical result of visual neuroscience is that cells in V1 are roughly divided into simple cells and complex cells, with simple cells forming linear filters sensitive to local features and complex cells pooling groups of simple cell response to form invariant representations. However the original sparse coding only models the simple cells [12, 14] without accounting for complex cells.

Group sparse coding extends sparse coding to include complex cells. Concretely, for each group j in the group sparse coding model, we assume that there is a downstream neuron (a complex cell) pooling the energy of the group $\|\mathbf{a_j}\|_2^2$. Interpreted this way, the group sparse coding model shares the same computational elements as the standard *energy model* [136] of complex cells. Similar to the energy model, response $\mathbf{a}$ from a group of simple cells with similar orientation selectivities are squared and summed to form inputs to a complex cell (the downstream neuron). Different from the energy model, the group sparse coding model further adds a square-root nonlinearity and enforces the response of complex cells to be sparse (Fig. 34; compare with Fig. 1C in [136]). With these additional structures, the group sparse coding model learns simple and complex cell receptive fields automatically from natural images, as opposed to the energy model where receptive fields are pre-specified. Note that the group sparse coding circuit structure in Fig. 34 is very similar to the independent subspace analysis (ISA) [150] – an extension of independent component analysis that was used to model complex cells.

The sparsity in the complex cell response induces decorrelation. Indeed, the model

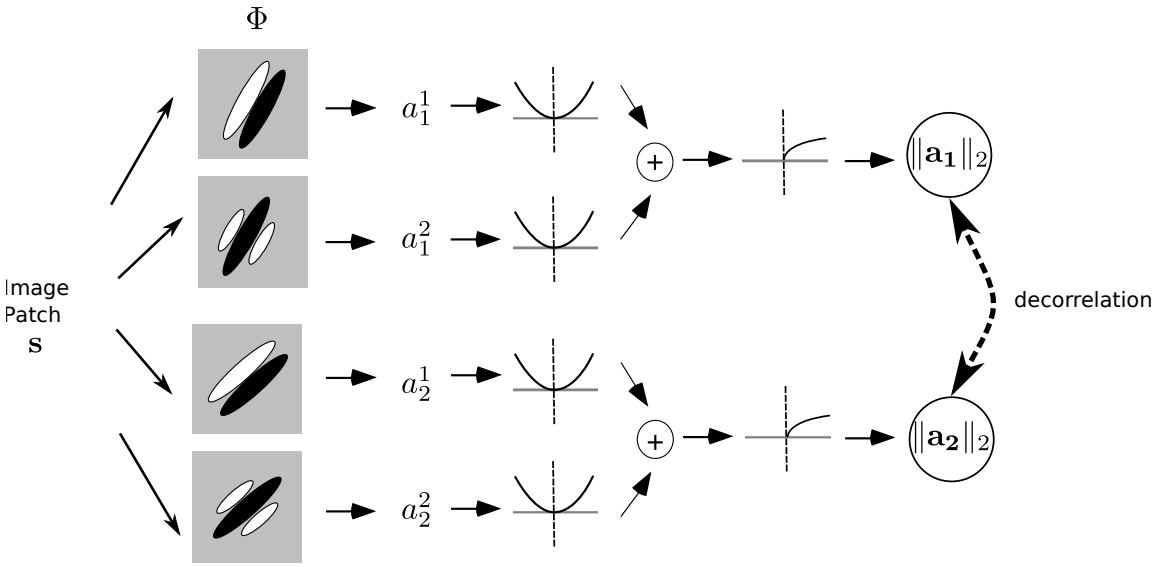


Figure 34: A stylized network implementation of group sparse coding. Viewed this way, group sparse coding model is a form of energy model that encourages decorrelation between complex cells. Decorrelation between complex cells can be achieved by recurrent connections and thresholding between the simple cells (see the main text for details).

complex cells are decorrelated (Fig. 35), similar to the model simple cells in the original sparse coding model (Fig. 31a).

The sparse inference/decorrelation process can be implemented in a dynamical system identical to the one introduced in Eq. 3 with an alternative thresholding function [49]:

$$\mathbf{a}^j = \widetilde{T}_\lambda(\mathbf{u}^j) = \begin{cases} 0 & \|\mathbf{u}^j\|_2 \leq \lambda \\ \mathbf{u}^j \left(1 - \frac{\lambda}{\|\mathbf{u}^j\|_2}\right) & \|\mathbf{u}^j\|_2 > \lambda \end{cases}. \tag{30}$$

In its current form however, it is unclear how the thresholding function could be implemented biologically: the threshold level depends on the internal state variables $\mathbf{u}$ (interpreted as the membrane potentials) from other neurons in the same group. This is not biologically feasible because membrane potentials are not directly observable from outside of the neuron. It remains a future direction how to implement group sparsity in a biologically feasible fashion.

An important advantage of group sparse coding compared to the original sparse coding

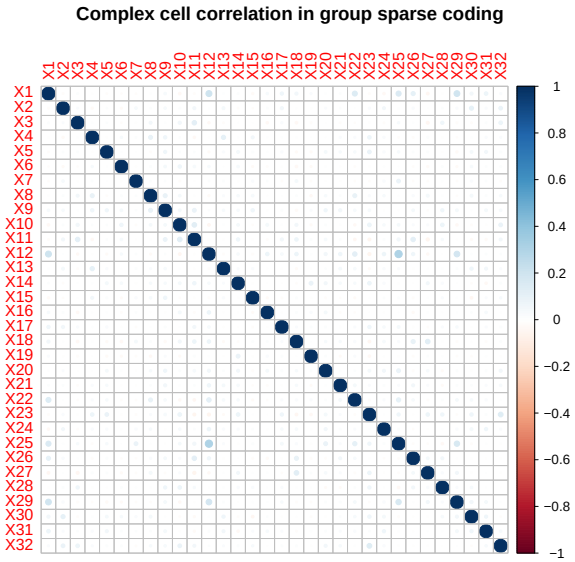


Figure 35: Correlation matrix of the group energy indicates that the model group response is decorrelated

is that it extracts invariant features. The complex cell response in the model is the aggregation of simple cells with similar orientation selectivity and different phase preference. As a result, the complex cells in the model are phase-invariant and orientation selective (Fig. 36; compare to similar properties in independent subspace analysis (ISA) [150]). This invariance is likely crucial in visual tasks such as recognition. Indeed, successive pooling and decorrelation have been shown to be particularly important for achieving good recognition performance in multi-layer deep convolutional networks [151].

4.7.3 Group sparsity prior and Frobenius-norm regularizer can be combined

The Frobenius-norm regularizer and the group sparsity prior are easily combined in a single model:

$$\{\hat{\Phi}, \hat{\mathbf{a}}\} = \arg\min_{\{\Phi,\mathbf{a}\}}(\|\mathbf{s} - \Phi\mathbf{a}\|_2^2 + \lambda \sum_j \|\mathbf{a^j}\|_2 + \gamma\|\Phi\|_F^2) \tag{31}$$

The resulting dictionary is a compact representation with a small number of effective groups (Fig. 37). We expect this model to exhibit both heterogeneity in the variance distribution as well as clustered correlation.

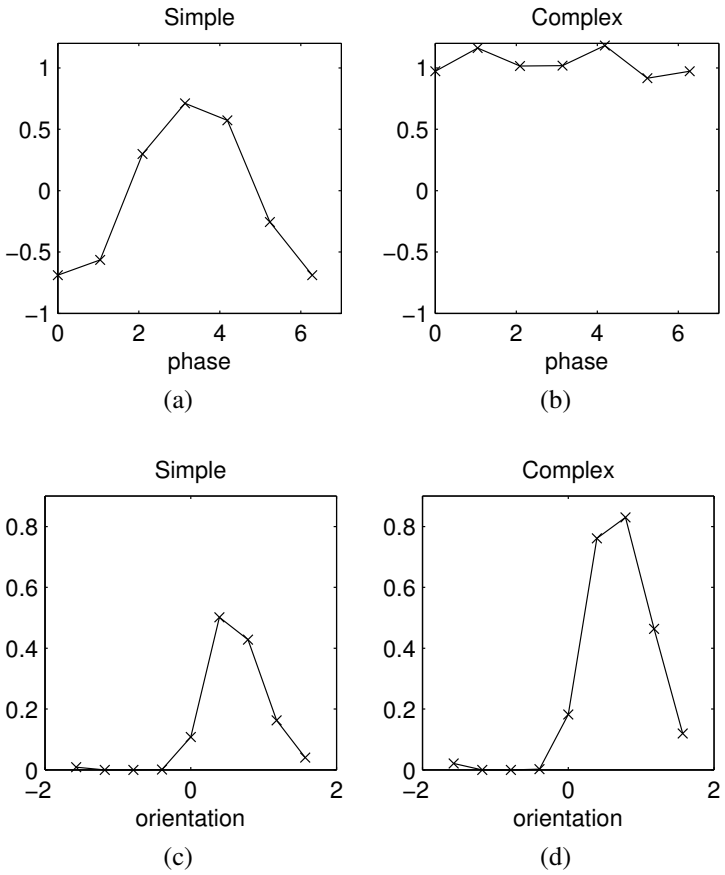


Figure 36: Comparison of the feature selectivity a simple cell and the corresponding complex cell in the group sparse coding model. (a) An example model simple cell is selective to the phase of a series of Gabor stimuli fitted to the RF of the cell. (b) The same set of stimuli induce similar response in the model complex cell, demonstrating phase-invariant response. (c) and (d) The model simple cell and complex cell have similar orientation tunings.

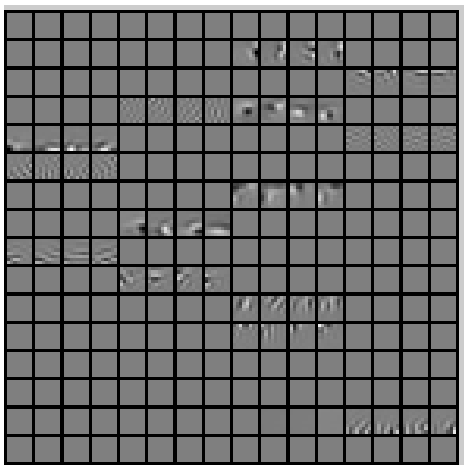


Figure 37: 4x overcomplete dictionary of size 256 learned with a Frobenius-norm regularizer and a group sparse prior of group size 4

4.7.4 Future works

It remains a question how the invariance in group sparsity contributes to visual tasks. Given the similarity between group sparse coding and ISA, which has proven to be successful in recognition tasks [152], it is likely that group sparse coding can provide similar computational benefits in visual tasks. Compared to ISA, group sparse coding is expected to learn more robust nonlinear features thanks to an overcomplete code.

In this study, we have treated all recorded neurons as simple cells with linear receptive fields. However, some of these cells may be complex, meaning that their RFs cannot be adequately described as linear filters. Indeed, an energy model gives a better prediction of the PSTH for many neurons than a linear receptive field model (results not shown). Further study is needed to establish the identity of each cell we studied and confirm that the complex cells are indeed decorrelated as predicted by the group sparse coding model.

The decorrelation we predict for complex cells is likely too strong compared to the real population. This is because the grouping in our model is fixed while the grouping boundaries in the experimental data appear far less rigid (Fig. 29a). A potential remedy is to introduce overlaps in the group structure. More advanced techniques need to be employed for inference in this kind of model [153]. With overlapping groups, it is expected that a

topographical organization similar to that observed in cats and primate would emerge [154, 155].

4.7.4.1 Additional computational benefits of group sparse coding

Group sparse coding may introduce additional computational benefits on top of invariance. For example, [156] demonstrated that when the signal has a group sparse structure, the group sparse coding is superior to the original sparse coding. Do natural images have group sparsity structure? Previous results on classification [157] and compressed sensing [155] suggest that this is indeed so. To test directly whether group sparsity is a better prior for natural images, we compare the correlation of the response between the sparse coefficients (Fig. 31a) and between the energies in the group sparse coding model (Fig. 35). A lower correlation would suggest a better match to the independence prior, thus implying a better statistical model. Preliminary results indicate that the mean correlation is slightly larger in the group sparse coding model, although correlations in both models are so small that this difference may prove to be insignificant. Dependencies between variances may need to be studied to further understand whether group sparsity prior is better for natural images.

4.7.5 Conclusion

To summarize, in this section we have shown that the group sparse coding model can better account for the correlation structure in the experimental data and can explain the emergence of complex cells in V1. The results suggest that sparse coding may be a strategy employed by higher stages in the visual pathway in addition to the simple cells at the first stage of visual processing.

4.8 Methods

4.8.1 Experimental methods

The visual stimulus, recording, and spike sorting methods were detailed in [158,159]. Here we give a brief summary.

4.8.1.1 Recording[4]

The recording was collected with a single shank, 32-channel planar silicon polytrode in anesthetized cats V1 through all cortical layers under visual stimulation. The polytrode has a channel spacing of 50μm, contact diameter of 23μm and thickness of 15μm.

The animals were anesthetized with an intramuscular injection of katamine (12 mg/kg) and acepromazine (0.3 mg/kg). Anesthesia was maintained with halothane (0.6 - 2 %) in a mixutre of nitrous oxide and oxygen (2:1). The spikes data were obtained by bandpassing the raw recorded electrical signal with a 4-th order Butterworth filter with a passband between 0.5kHz and 10kHz. Single unit activities were extracted by a standard tetrode spike sorting procedure where similar spikes were clustered with k-means followed by manual cleanup[5]. The neural response was then defined as the instantaneous spike count binned at 10Hz.

4.8.1.2 Stimulus presentation in the experiment

We analyze single trial as well as repeated trials population response to full field natural movies stimulation. The movies were recorded with a digital camera with the field of view matched to the visual angle at which the movies were later presented to the animal. The movies were recorded at 300Hz frame rate and subsequently resampled to 150Hz. The movies used in this study include three long movies “duck8”, “duck30”, “cat”, and a short 30s repeated movie clip taken from the “duck8” movie.

4.8.1.3 STRF estimation[6]

In the correlation study, the spatial temporal receptive field was estimated using spike-triggered average of the response to the natural movie stimulus.

[4]The recording was conducted in Dr. Charles Gray’s lab at Montana State University.
[5]Spike sorting was carried out by Dr. Urs Köster.
[6]The STRFs were estimated by Dr. Ian Stevenson.

4.8.1.4 *Mitigating the effect of anesthesia*

We took care to only analyze subsets of the data where the effect of anesthesia is minimum. For example among the 60 repeated trials in the original experiment, only the latter 40 trials were used to calculate the PSTH because the LFP of the first 20 trials suggests a different brain state. Specifically, the LFP of the first 20 trials shows a strong low-frequency oscillation that is likely a consequence of anesthesia. Another symptom of the low-frequency oscillation is that the correlation between cells in the first 20 trials was not stationary (Sect. 6.6).

4.8.2 Modeling methods

4.8.2.1 *Stimulus presentation to the models*

Unless otherwise noted, the results presented in this chapter were computed with a 200s single trial movie segment from the movie "duck8". To match the bin size (100 ms) used to calculate the instantaneous spike count in the experiment, in the simulation we used frames 100 ms apart. While this ignores some high-frequency dynamics in the stimulus, temporal dynamics is not the focus of this study.

To prepare the raw movie stimulus for simulation, we prepared the movies in a similar manner as described before [158]. Specifically we downsampled the raw movie frames by a factor of 16 to 32×24 pixels, whitened and cropped out 8×8 frame centers to match the location and the spatial extent of the classical receptive fields [158].

4.8.2.2 *Dictionary learning*[7]

To adapt the sparse coding models to the statistics of the natural movies, 1800 downsampled and whitened 32×24 frames each 0.6s apart were selected from the duck8, duck30, and cat movies. 8×8 image patches were then sampled randomly from these frames and presented to the model. The low-frequency dictionary in Fig. 28c was learned from patches directly cropped from the movie frames without downsampling and whitening. These local patches

[7]The base code (without the stimulus preparation) for learning sparse coding dictionary was provided by Dr. Adam Charles.

only have the low-frequency components due to the limited resolution of the raw movie frames.

The learned dictionary elements were taken as the classical receptive fields. This is justified because the dictionary elements are approximately the same as the feedforward classical receptive fields mapped out using sparse dots [12].

4.8.2.3 Inference

To infer the sparse coefficients in sparse coding models with the Laplacian prior and with the group sparsity prior, fast iterative shrinkage-thresholding algorithm (FISTA) algorithm implemented in the SPAse Modeling Software (SPAMS) package[8] was used. Using this implementation (as opposed to l1ls or LCA) enables us to learn large (e.g. 16x overcomplete) dictionaries in hours.

In all inference, the trade-off parameter $\lambda = 0.2$ and the Frobenius-norm trade-off parameter $\gamma = 0.02$. With this parameter choice, all inference has an rMSE less than 15%.

4.8.2.4 Measuring correlation

Correlation between cells is measured by Pearson's correlation between the spike count vectors. This correlation measure is sensitive to the choice of bin size. Using a small bin size results in artificially small correlation due to noise/spike jittering [160]. Indeed in our dataset using bins synchronized to the frame rate (150Hz) produces correlations close to zero. To reduce the effect of noise, we measure the correlation using 100ms bins – a size within the range of correlation timescale in the brain [160]. Using 250ms bins leads to similar results.

The receptive field similarity was measured using cosine similarity (the uncentered Pearson's correlation). The similarity was measured between the STRFs in the experiment and the spatial RFs in the models since there is no temporal component in the learned dictionary.

[8] `http://spams-devel.gforge.inria.fr/`

To measure the similarities between 2D distributions from models and from experiments (e.g. between Fig. 30b and Fig. 29c), we first estimated the probabilistic densities through Gaussian kernel smoothing, then measured the statistical distance by integrating the squared difference between the densities. We confirmed through simulations on the Gaussian distributions that this measure is close to 0 when the two samples come from the same distribution and it grows with increasing difference between the distributions. This measure is advantageous to other statistical similarity measures such as the KL-divergence in this case because: first, it is easy to apply in 2D; second, it is not sensitive to the difference in the number of samples; third, it gives a meaningful measure even when the two distributions are very different (KL tends to have numerical issues in this case due to dividing by 0 probability).

4.8.2.5 Measuring invariance

To quantify the phase-invariant property of the model complex cell, we first fitted a Gabor patch to the RF of a corresponding simple cell that feeds into the complex cell. We then varied the phase of the Gabor patch and measured the response from both the simple cell and the complex cell (the square-rooted total group energy).

4.8.2.6 Software

Raw spike data were pre-processed in Python with the neo package[9]. Simulated electrophysiology was conducted in Matlab®. In the invariance experiment, Gabor filters were fitted to the RF using the "Auto Gaussian & Gabor fits" package[10]. Most of the statistical analysis and visualization was done in R. In particular, we used the "kde.test" function in the "ks" package for kernel smoothing and for measuring the statistical distance between smoothed 2D densities.

[9] `https://pythonhosted.org/neo/`

[10] `http://www.mathworks.com/matlabcentral/fileexchange/31485-auto-gaussian---gabor-fits`

CHAPTER V

CONCLUSION

5.1 Contributions

In this dissertation, we have demonstrated that sparse coding and its variants are consistent with and in many cases can account for key aspects of neural response in the primary visual cortex, including contextual and nonlinear effects, inhibitory interneuron properties, as well as variance and correlation distributions of population response. There are three major contributions:

1. Biologically, the results demonstrate that detailed physiological properties can be understood through the constraints and coding goals of the sensory system. In particular, the results illustrate a number of biological constraints crucial for understanding population neural response, including the response sparsity (Chap. 2), number of cells (Chap. 3), synaptic weights, and hierarchy (Chap. 4).

2. In terms of modeling, the results bring theoretical neural coding models one step closer to biological reality. These biological plausible models enable direct comparison with experimental results and make predictions that can be verified by further experiments.

3. Computationally, the results presented here suggest new learning and inference models that have unique advantages over the original sparse coding (Chap. 4). In particular, these models learn features that are more concise and invariant.

5.2 Future works

The present work can be extended in several directions.

First of all, in the current study, diverse physiological phenomena are investigated separately with different variations of the sparse coding model. It is important to understand how compatible these models are with each other. For example, do complex cells in the group sparse coding model (Chap. 4) exhibit non-classical receptive field effects (Chap. 2) (similar to those studied in [67])?

Currently our models treat the cortical population as a homogeneous population with no laminar structures. However, different layers in the cortex likely have different circuit structures, cell species, and response properties. For example, it was found that simple cells dominate layer 4 and 6 in V1 while complex cells distribute across all layers in cats [161][1]. A more complete coding model needs to capture this kind of laminar specialization. For instance, a more general model may consist of the original sparse coding model which is specific to simple cells and is likely useful for modeling layer 4 response, and the group sparse coding model which is more useful for other layers. How to align these different models to the layered structure is an important future topic.

In many respects the feature learning perspective adopted in this dissertation parallels recent advancement in computer vision where deep networks with features adapted to large numbers of natural images outperform approaches using hand-crafted features by a large margin and achieve state-of-the-art results [151]. In certain tasks, these algorithms even start to approach biological performance [162]. Despite this success, current computer vision algorithms are still no match for the generality and efficiency of human vision. The results in this dissertation elucidate certain computational primitives potentially employed by biological neural networks and may have implications for neural network models used in computer vision. How to transfer the knowledge learned from biology to artificial systems is an important future direction.

[1] It is unclear whether this applies to rodents [147].

CHAPTER VI

APPENDICES

6.1 Effects of changing simulation parameters in the nCRF study

As mentioned in Chap. 2, changing parameters such as the sparsity level λ and the simulation time changes the quantitative results for the simulation. For example, when simulating the dynamical system all the way to a steady state response (1000 integration time steps), the surround suppression index distribution shows an unrealistically large number of cells that demonstrate essentially complete suppression (Fig. 38; compare to Fig. 4b).

While such parameter changes can sometimes show biophysically unrealistic responses (especially for population statistics), these parameter changes can sometimes account for apparently conflicting reports in the physiology literature. For example, Fig. 39 shows the surround suppression index produced by the model when we lowered the tradeoff parameter λ (i.e., allowing more simultaneously active cells) and increased the number of integration time steps for the system (i.e., presenting the stimulus to the system for a longer time, thereby letting the network converge more fully). In this case, the model produced a surround suppression index that had most cells being suppressive, peaking near a value of 70%. Interestingly, this is qualitatively similar to other reports from the physiology literature [15, 81, 124] that are apparently conflicting with the more typical reports of most cells being non-suppressive, perhaps due to a different experimental preparation.

For another example of a case when a different set of parameters could be used to replicate an alternative report from the literature, consider the fact that some neurons in cat and macaque V1 actually show facilitation with iso-oriented surround stimuli when using low center contrast stimuli [16, 163] (Fig. 40a; not observable in Fig. 8d). Such effects were previously modeled mechanistically [163–165] as a result of changing balance between the excitation and inhibition under different contrast (input strengths). We show that

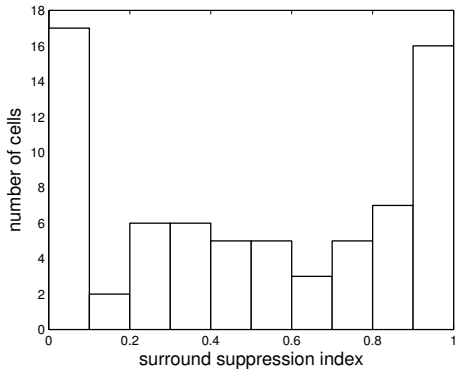


Figure 38: Surround suppression index distribution under a different parameter setting. Related to Fig. 4b. With steady-state response of the model and otherwise default parameters, the surround suppression index distribution shows physiologically unrealistic large percentage of cells with complete suppression.

this facilitation effect also emerges from our functional model response when the tradeoff parameter is set to $\lambda = 0.1$ and all other stimulation procedures are kept the same (see Fig. 40). This might be understood as a result of the nonlinear change of competition between neurons with different contrast/input drives (potentially with secondary effects such as disinhibition). Related to our result, Coen-Cagli et al. [66] recently showed that surround facilitation could arise under certain stimulus conditions in a statistical model adapted to natural scenes. In their model, how much the center response is normalized (modulated) by the surround response is dependent on the relative stimulus contrast in the center and surround, as well as the receptive field correlations between the center and surround (similar to the present model, see Sect. 2.3 for more details on the differences). We note that contrast-dependent facilitation might be beneficial for perceptual tasks such as contour integration. For example, when the center contrast is low, it is perceptually helpful to enhance the response to fill in the gap and complete the contour. However, when the center contrast is high, it is more efficient to keep the response in check and relegate some of the responsibility for the representation to other cells.

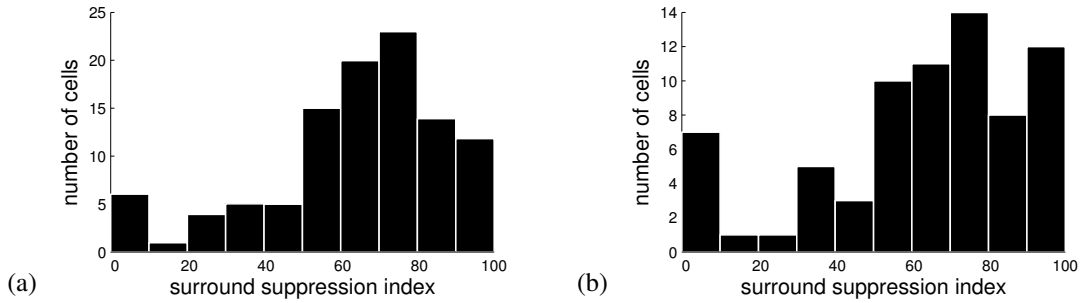


Figure 39: Surround suppression index distribution under another parameter setting. Related to Fig. 4b. (a) Physiologically measured index from an experiment on macaque monkeys (N=105); data replotted from [15, Figure 2C]; (b) Simulation of the surround suppression index distribution with lower sparsity and longer convergence times ($\lambda = 0.05$ and 1000 integration time steps). Note that the majority of neurons are surround suppressive in this case.

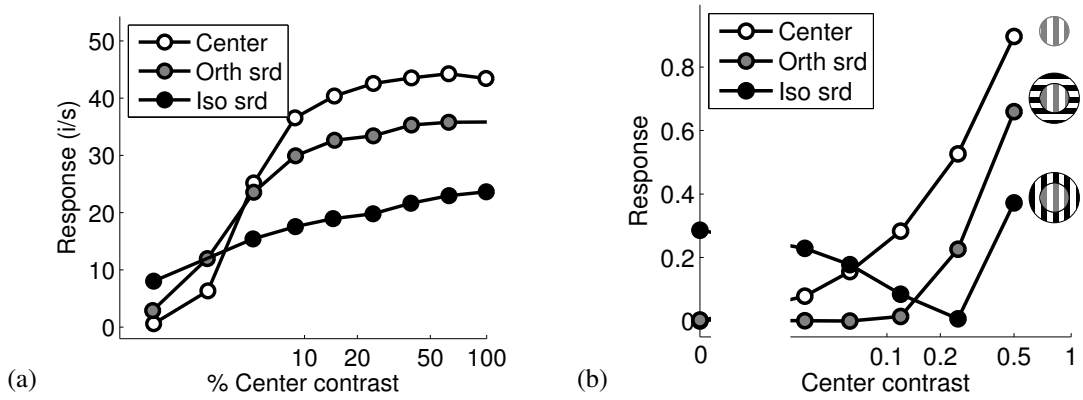


Figure 40: Facilitatory influence. Related to Fig. 8. (a) Facilitatory influence from the iso-surround at low center contrast observed in cats; data replotted from [16, Figure 5]; (b) A simulated neuron demonstrates a similar effect when the tradeoff parameter is set to $\lambda = 0.1$.

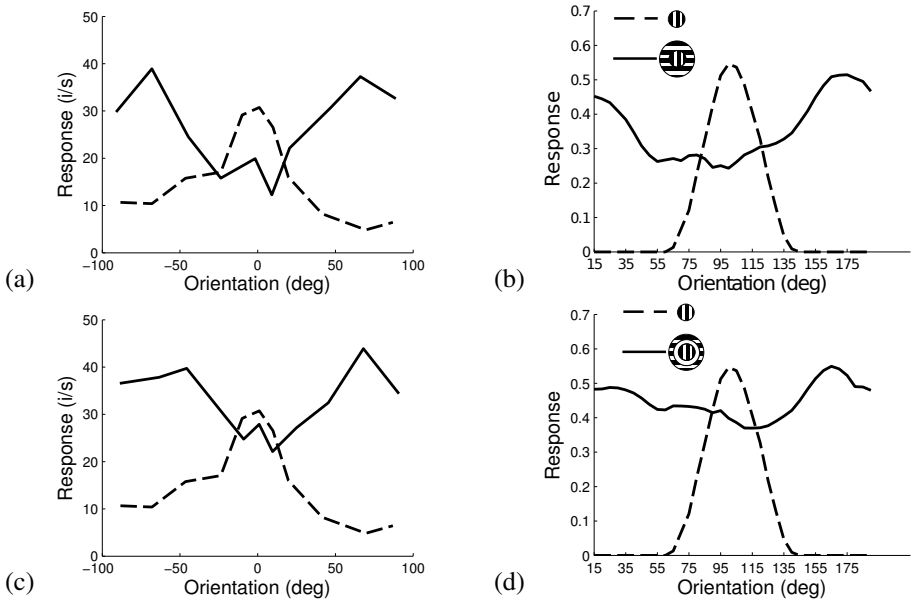


Figure 41: Spatial organization of surround orientation tuning. Orientation tuning with "gap" in between center and surround. (a) Physiology without gap; data replotted from [4, Figure 4D]; (b) Simulation without gap; (c) Physiology with gap; data replotted from [4, Figure 4E]; (d) Simulation with gap. Parameters same as in Fig. 7.

6.2 Other Miscellaneous nCRF Effects

Using a similar setup as in Fig. 7, the suppressive effect of the surround is smaller if there is a gap separating it from the center (Fig. 41). The extent of contextual effect is therefore modulated by the area the surround is covering. The larger the surround, the stronger the effect.

6.3 Mathematical derivations of model receptive fields

To study the orientation tuning of cells in the inhibitory interneuron model (Chap. 3) , we map the receptive field using sparse spots of light as in [12]. In the following derivation, without loss of generality we consider variables that could be both positive and negative instead of constraining the signs as in Chap. 3.

6.3.1 RFs of excitatory cells

Here we derive the RFs of principal cells in a linear generative model. The stimuli array of sparse dots used for mapping RF is represented by an identity matrix $I = [\mathbf{s}_1, ..., \mathbf{s}_M]$, and

the rows of the response matrix A are the RFs.[1] In a linear generative model:

$$I \approx \Phi A. \tag{32}$$

First consider the case when Φ (and thus $\Phi^T\Phi$) is full rank. The receptive field mapped with sparse dots is the Moore-Penrose inverse of the dictionary matrix:

$$A \approx \Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T. \tag{33}$$

Specifically when Φ is orthogonal,

$$A \approx \Phi^T, \tag{34}$$

indicating that the RFs of the excitatory cells are approximately the same as the dictionary elements.

For an arbitrary overcomplete Φ, there is no unique solution to Eq. (32). If one regularizes the problem with a sparsity constraint on the columns of A, A can be inferred using a linear dynamical system coupled with a nonlinearity (Eq. (8)) [23]. The solution at steady state under sparse dots stimuli, ignoring the thresholding ($\mathbf{a} \approx \mathbf{u}$) can be written as:

$$A \approx \Phi^T I - (\Phi^T\Phi - I)A, \tag{35}$$

which implies

$$\Phi^T\Phi A \approx \Phi^T, \tag{36}$$

similar to Eq. (33). When the dictionary elements are sufficiently different from one another (i.e. $\Phi^T\Phi$ is close to identity), again we have Eq. (34). Empirically in simulations we observe that the tuning property of the RFs are indeed similar to that of the dictionary when using an overcomplete dictionary learned from natural scenes. A similar observation was also made previously in [12].

[1] Here we mix the signs in the representation for the sake of brevity. In simulations we separate out the signs: we map the ON part of the RF with positive-valued dot, and the OFF part by the negative-valued.

6.3.2 RFs of inhibitory cells in the direct implementation

When the Gramian matrix is decomposed directly (Eq. (12)), the RFs of the inhibitory cells are $(\Phi^T\Phi)_+A$. According to Eq. (36), this is approximately the same as the dictionary elements (orientation tuned).

6.3.3 RFs of inhibitory cells in the Gramian decomposition

When the Gramian matrix $\Phi^T\Phi$ is used to represent the recurrent inhibition directly (Fig. 16b), the RFs of the interneurons are columns of ΦA, where columns of A are the responses of principal cells to sparse dots stimuli. It is straightforward to infer from Eq. (32) that the RFs of the interneurons are dots (simulation in Fig. 16b) since ΦA is the identity matrix.

6.3.4 RFs of inhibitory cells in low-rank decomposition

Denote the SVD of Φ as

$$\Phi = \Omega\Lambda\Upsilon^T, \tag{37}$$

so that

$$\Phi^T\Phi = \Upsilon\Lambda\Omega^T\Omega\Lambda\Upsilon^T = \Upsilon\Lambda^2\Upsilon^T. \tag{38}$$

From Eq. (17), since S is sparse, $\Phi^T\Phi \approx L = U\Sigma V^T$. Therefore $U \approx \Upsilon \approx V$ and $\Sigma \approx \Lambda^2$ up to scaling constants. The RFs of the inhibitory neurons implementing the low-rank part of the connectivity matrix can thus be approximated up to a scaling constant by:

$$V^TA \approx \Upsilon^TA = (\Lambda^{-1}\Omega^T\Omega\Lambda)\Upsilon^TA = (\Omega\Lambda^{-1})^T(\Omega\Lambda\Upsilon^TA) = (\Omega\Lambda^{-1})^T(\Phi A) \overset{(32)}{\approx} (\Omega\Lambda^{-1})^T.$$

In words, this means that the RFs of these interneurons can be approximated by the (weighted) singular vectors of the dictionary Φ. Since the dictionary is learned from natural scenes, the singular vectors resemble the PCA components learned from natural scenes, which are known to be sinusoidal [166]. The simulation confirms this (Fig. 19a).

6.4 Feedforward inhibition

While Chap. 3 addressed the role of inhibition in the recurrent cortical connections, similar considerations arise when modeling the feedforward inhibition prevalent in the thalamocortical system [92]. The classical push-pull feedforward model [167] proposed the existence of tuned feedforward inhibitory neurons with RFs that "mirror" those of the excitatory neurons and supply disynaptic hyper-polarization to the principal cells in layer 4. To account for the discrepancy in the numbers of inhibitory and excitatory cells, it was suggested that these inhibitory neurons are multiplexed (i.e. a single inhibitory neuron supplies inputs to multiple excitatory neurons [52]), though there is uncertainty in the potential implementations of this multiplexing. Assuming that the interneurons mediating the feedforward inhibition is a separate population from the recurrent inhibition population, we show that multiplexing can be implemented naturally when the representation is overcomplete (Fig. 42). Specifically we rewrite the feedforward transformation (Φ^T) in Eq. (20) as follows:

$$\Phi^T = \Phi_+^T + \Phi_-^T = \Phi_+^T + \Psi\Omega.$$

We can interpret Ω as the connectivity matrix from the inputs to the feedforward inhibitory interneurons and Ψ as the connectivity matrix from the interneurons to the principal cells. Note that the dimension of Ω can be arbitrarily assigned as long as the decomposition holds. A simple choice is to make Ω a full-rank dictionary matrix ($N \times N$) with Gabor-like elements. As a result when the principal cell representation is overcomplete ($\Psi : M \times N$ with $M > N$), fewer feedforward interneurons than principal cells are needed. Furthermore, the RFs of the feedforward interneurons are Gabor-like, similar to those in the classical push-pull model [52]. Note that Ψ is uniquely defined given Φ and Ω, because in this case Ω^{-1} exists .

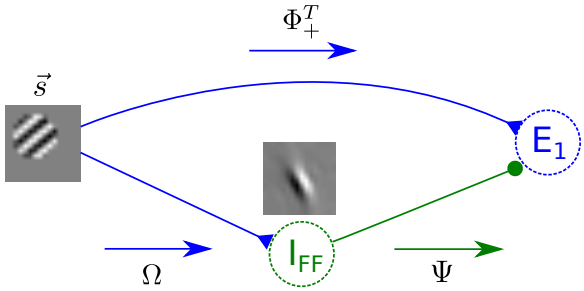


Figure 42: Feed-forward inhibition. Feedforward push-pull could also be implemented with fewer inhibitory neurons than excitatory neurons.

6.5 Global inhibition

While many implementations of inhibitory interneurons are discussed in Chap. 3, an alternative strategy to improve the E/I ratio is to use a global inhibition approach. In the network of Eq. (8), we can offset the baseline excitatory synaptic weights by a constant $c > 1$ to make all of the individual synaptic weights positive: $(c - \langle \phi_j, \phi_i \rangle) > 0, \forall j$. With this offset to the individual weights, the equivalent dynamics model becomes:

$$\tau \dot{u}_i(t) = \langle \phi_i, \mathbf{s} \rangle + \underbrace{\sum_{j \neq m} (c - \langle \phi_j, \phi_i \rangle) a_j(t)}_{\text{Excitatory}} - \underbrace{c \sum_{j \neq m} a_j(t)}_{\text{Global Inhibitory}} - u_i(t) \tag{39}$$

where we now have a single non-orientation tuned inhibitory cell (Fig. 43) that sums over the activity of all principal cells. In biological reality global inhibition is unlikely to manifest in this manner, if at all. A more likely scenario is that this type of interneuron inhibits only a local population of principal cells. This may in fact contribute to diverse orientation tunings of the inhibitory neurons, depending on the local distribution of principal cell orientation tuning properties (e.g. location on the pinwheel).

We note several additional features of this global inhibition implementation of sparse coding. First, the global inhibition rises and falls with the overall excitatory activity in the network. This implies that the model inhibitory neurons have inherently higher firing rates, reminiscent of one class of inhibitory neurons in the cortex that are fast-spiking [130]. Second, the network equation indicates that excitatory cells only connect to those sharing overlapping RFs, reminiscent of the connectivity pattern seen in layer 2/3 where cells with

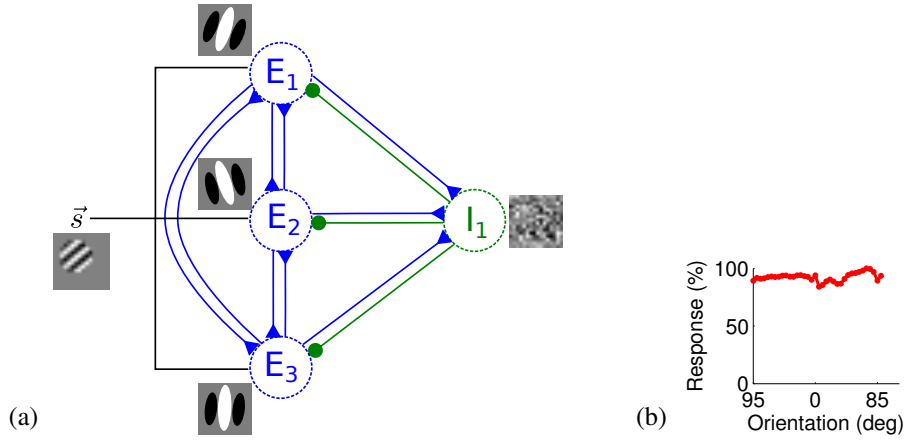


Figure 43: Global inhibition. (a) The recurrent network that implements the global inhibition (Eq. (39)). I_1 pools all activities from the excitatory population, weighs them by c, and projects back to the excitatory population. (b) The orientation tuning curve of the inhibitory neuron I_1.

overlapping inputs share connections [145, 146], while inhibitory interneurons form dense connections with all neighboring excitatory neurons, similar to the pattern seen in physiology [77]. Third, the excitatory synaptic weights may result from combined Hebbian and anti-Hebbian learning with rates that depend on the overlap of the RFs. Functionally, the connection pattern in Eq. (39) can be interpreted as a way for excitatory connections to share responsibilities for representing inhibition by shifting the baseline activity to a non-zero value. The cost of such an implementation is that the excitatory population has a higher gain, which may potentially induce a higher sensitivity to noise.

6.6 Unstationary correlation in some experiment trials

The correlation between cells is not stationary for the first 20 trials of the short movie experiment (Fig. 44) potentially due to anesthesia.

6.7 Response vs. receptive field correlation in additional data sets[2]

See Fig. 45. This confirmed the shape of the distribution in Fig. 29c.

[2]Data provided by Dr. Ian Stevenson

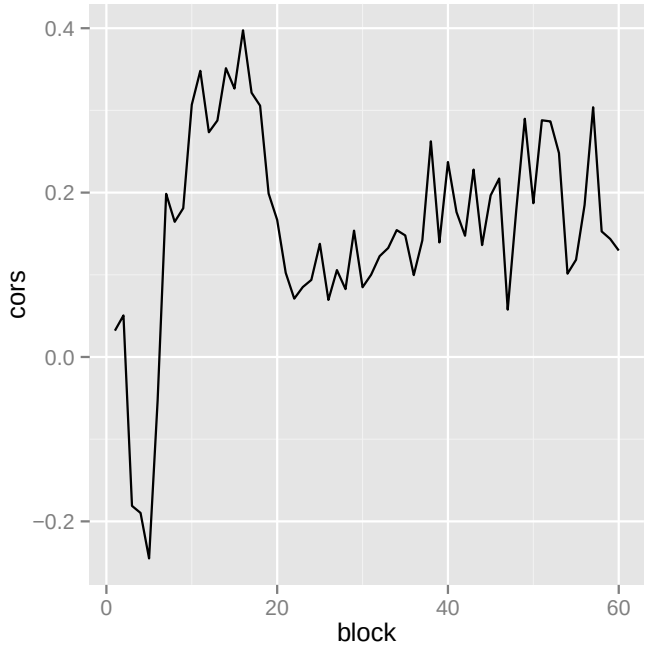


Figure 44: The response correlation between two cells evolving over the 60 trials of repeated short movies.

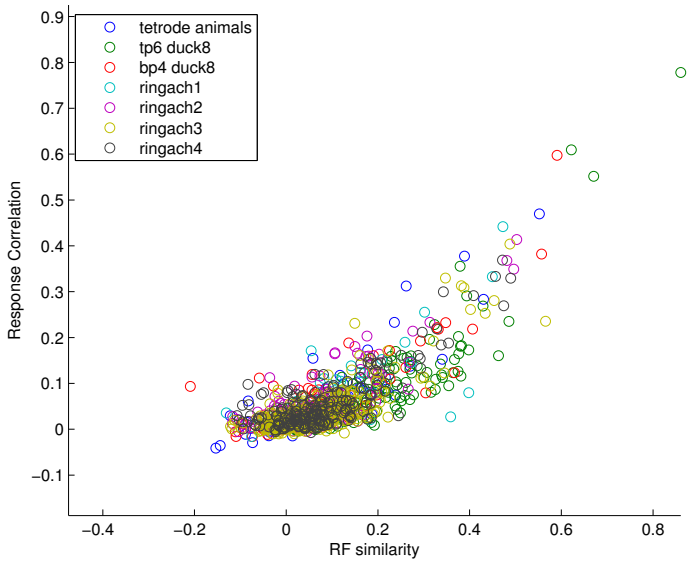


Figure 45: The response vs. RF correlation in a few other experiments in addition to the Beck P4 data we analyzed in the main text. Response sampled at 250ms.

REFERENCES

[1] D. Rose, "Responses of single units in cat visual cortex to moving bars of light as a function of bar length," *The Journal of Physiology*, vol. 271, no. 1, pp. 1–23, 1977.

[2] M. Sceniak, D. Ringach, M. Hawken, and R. Shapley, "Contrast's effect on spatial summation by macaque V1 neurons," *Nature Neuroscience*, vol. 2, pp. 733–739, 1999.

[3] G. Walker, I. Ohzawa, and R. Freeman, "Suppression outside the classical cortical receptive field.," *Visual neuroscience*, vol. 17, no. 3, pp. 369–379, 2000.

[4] H. Jones, W. Wang, and A. Sillito, "Spatial organization and magnitude of orientation contrast interactions in primate V1," *Journal of neurophysiology*, vol. 88, no. 5, pp. 2796–2808, 2002.

[5] B. Webb, N. Dhruv, S. Solomon, C. Tailby, and P. Lennie, "Early and late mechanisms of surround suppression in striate cortex of macaque," *Journal of Neuroscience*, vol. 25, no. 50, pp. 11666–11675, 2005.

[6] J. Cavanaugh, W. Bair, and J. Movshon, "Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons," *Journal of Neurophysiology*, vol. 88, no. 5, pp. 2547–2556, 2002.

[7] B. Skottun, A. Bradley, G. Sclar, I. Ohzawa, and R. Freeman, "The effects of contrast on visual orientation and spatial frequency discrimination: a comparison of single cells and behavior," *Journal of Neurophysiology*, vol. 57, no. 3, pp. 773–786, 1987.

[8] H. Alitto and W. Usrey, "Influence of contrast on orientation and temporal frequency tuning in ferret primary visual cortex," *Journal of neurophysiology*, vol. 91, no. 6, pp. 2797–2808, 2004.

[9] A. B. Bonds, "Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex," *Visual Neuroscience*, vol. 2, no. 01, pp. 41–55, 1989.

[10] T. Freeman, S. Durand, D. Kiper, and M. Carandini, "Suppression without inhibition in visual cortex," *Neuron*, vol. 35, no. 4, pp. 759–771, 2002.

[11] N. Priebe and D. Ferster, "Mechanisms underlying cross-orientation suppression in cat visual cortex," *Nature Neuroscience*, vol. 9, no. 4, pp. 552–561, 2006.

[12] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images.," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[13] J. A. Hirsch, L. M. Martinez, C. Pillai, J.-M. Alonso, Q. Wang, and F. T. Sommer, "Functionally distinct inhibitory neurons at the first stage of visual cortical processing," *Nat Neurosci*, vol. 6, pp. 1300–1308, Dec. 2003.

[14] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[15] H. Jones, K. Grieve, W. Wang, and A. Sillito, "Surround suppression in primate V1," *Journal of Neurophysiology*, vol. 86, no. 4, pp. 2011–2028, 2001.

[16] D. Somers, E. Todorov, A. Siapas, L. Toth, D. Kim, and M. Sur, "A local circuit approach to understanding integration of long-range inputs in primary visual cortex.," *Cerebral Cortex*, vol. 8, no. 3, pp. 204–217, 1998.

[17] B. Olshausen and D. Field, "How close are we to understanding V1?," *Neural Computation*, vol. 17, no. 8, pp. 1665–1699, 2005.

[18] D. Hubel and T. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.

[19] D. Hubel and T. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of Physiology*, vol. 195, no. 1, p. 215, 1968.

[20] B. Haider, M. Hausser, and M. Carandini, "Inhibition dominates sensory responses in the awake cortex," *Nature*, vol. 493, pp. 97–100, Jan. 2013.

[21] A. Wohrer, M. D. Humphries, and C. K. Machens, "Population-wide distributions of neural activity during perceptual decision-making," *Progress in Neurobiology*, vol. 103, no. 0, pp. 156 – 193, 2013. ¡ce:title¿Conversion of Sensory Signals into Perceptions, Memories and Decisions¡/ce:title¿.

[22] E. Simoncelli, "Vision and the statistics of the visual environment," *Current opinion in Neurobiology*, vol. 13, no. 2, pp. 144–149, 2003.

[23] C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, "Sparse coding via thresholding and local competition in neural circuits," *Neural Computation*, vol. 20, no. 10, pp. 2526–2563, 2008.

[24] M. Zhu and C. J. Rozell, "Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system," *PLoS Comput Biol*, vol. 9, p. e1003191, 08 2013.

[25] H. Hartline, "The response of single optic nerve fibers of the vertebrate eye to illumination of the retina.," *American Journal of Physiology*, vol. 121, pp. 400–415, 1938.

[26] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[27] P. Seriès, J. Lorenceau, and Y. Frégnac, "The "silent" surround of V1 receptive fields: theory and experiments," *Journal of physiology-Paris*, vol. 97, no. 4-6, pp. 453–474, 2003.

[28] M. Carandini, D. Heeger, and J. Movshon, "Linearity and gain control in V1 simple cells," *Cerebral Cortex: Models of Cortical Circuits*, pp. 401–444, 1999.

[29] W. Vinje and J. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[30] W. Vinje and J. Gallant, "Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1," *Journal of Neuroscience*, vol. 22, no. 7, pp. 2904–2915, 2002.

[31] G. Felsen, J. Touryan, and Y. Dan, "Contextual modulation of orientation tuning contributes to efficient processing of natural stimuli," *Network: Computation in Neural Systems*, vol. 16, no. 2-3, pp. 139–149, 2005.

[32] B. Haider, M. Krause, A. Duque, Y. Yu, J. Touryan, J. Mazer, and D. McCormick, "Synaptic and Network Mechanisms of Sparse and Reliable Visual Cortical Activity during Nonclassical Receptive Field Stimulation," *Neuron*, vol. 65, no. 1, pp. 107–121, 2010.

[33] T. Albright and G. Stoner, "Contextual Influences on Visual Processing," *Annual Review of Neuroscience*, vol. 25, no. 1, pp. 339–379, 2002.

[34] N. J. Priebe and D. Ferster, "Mechanisms of neuronal computation in mammalian visual cortex," *Neuron*, vol. 75, no. 2, pp. 194–208, 2012.

[35] A. Dobbins, S. W. Zucker, and M. S. Cynader, "Endstopped neurons in the visual cortex as a substrate for calculating curvature," *Nature*, vol. 329, pp. 438–441, Oct. 1987.

[36] D. Field and A. Hayes, "Contour integration and the lateral connections of V1 neurons," in *The Visual Neurosciences* (L. Chalupa and J. Werner, eds.), pp. 1069–1079, Cambridge, MA: The MIT Press, 2004.

[37] V. A. F. Lamme, "Beyond the classical receptive field: contextual modulation of V1 responses," in *The Visual Neurosciences* (L. Chalupa and J. Werner, eds.), pp. 720–732, Cambridge, MA: The MIT Press, 2004.

[38] M. Rehn and F. Sommer, "A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields," *Journal of Computational Neuroscience*, vol. 22, no. 2, pp. 135–146, 2007.

[39] B. Olshausen and D. Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[40] J. E. Niven and S. B. Laughlin, "Energy limitation as a selective pressure on the evolution of sensory systems," *Journal of Experimental Biology*, vol. 211, no. 11, pp. 1792–1804, 2008.

[41] E. Baum, J. Moody, and F. Wilczek, "Internal representations for associative memory," *Biological Cybernetics*, vol. 59, no. 4, pp. 217–228, 1988.

[42] H. L. Yap, A. S. Charles, and C. J. Rozell, "The restricted isometry property for echo state networks with applications to sequence memory capacity," in *Proceedings of IEEE Statistical Signal Processing Workshop*, pp. 580–583, IEEE, 2012.

[43] J. Wolfe, A. Houweling, and M. Brecht, "Sparse and powerful cortical spikes," *Current Opinion in Neurobiology*, pp. 306–312, 2010.

[44] R. N. S. Sachdev, M. R. Krause, and J. A. Mazer, "Surround suppression and sparse coding in visual and barrel cortices," *Frontiers in Neural Circuits*, vol. 6, no. 00043, p. doi: 10.3389/fncir.2012.00043, 2012.

[45] M. Spratling, "Predictive Coding as a Model of Response Properties in Cortical Area V1," *Journal of Neuroscience*, vol. 30, no. 9, pp. 3531–3543, 2010.

[46] L. Perrinet, "Role of homeostasis in learning sparse representations," *Neural computation*, vol. 22, no. 7, pp. 1812–1836, 2010.

[47] 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 Comput Biol*, vol. 7, p. e1002250, 10 2011.

[48] A. Balavoine, J. Romberg, and C. J. Rozell, "Convergence and rate analysis of neural networks for sparse approximation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 1377–1389, sept. 2012.

[49] A. S. Charles, P. Garrigues, and C. J. Rozell, "A common network architecture efficiently implements a variety of sparsity-based inference problems," *Neural Computation*, vol. 24, pp. 3317–3339, Sept. 2012.

[50] S. Shapero, A. Charles, C. J. Rozell, and P. Hasler, "Low power sparse approximation on reconfigurable analog hardware," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 2, pp. 530 –541, sept. 2012.

[51] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An Interior-Point Method for Large-Scale ℓ_1-Regularized Least Squares," *IEEE journal of selected topics in signal processing*, vol. 1, no. 4, pp. 606–617, 2007.

[52] J. Hirsch and L. Martinez, "Circuits that build visual cortical receptive fields," *Trends in neurosciences*, vol. 29, no. 1, pp. 30–39, 2006.

[53] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, pp. 801–808, 2007.

[54] C. Wang, C. Bardy, J. Y. Huang, T. FitzGibbon, and B. Dreher, "Contrast dependence of center and surround integration in primary visual cortex of the cat," *Journal of Vision*, vol. 9, no. 1, p. doi: 10.1167/9.1.20, 2009.

[55] X. Song and C. Li, "Contrast-dependent and contrast-independent spatial summation of primary visual cortical neurons of the cat," *Cerebral Cortex*, vol. 18, no. 2, pp. 331–336, 2008.

[56] J. Levitt and J. Lund, "Contrast dependence of contextual effects in primate visual cortex," *Nature*, vol. 387, no. 6628, pp. 73–76, 1997.

[57] M. Carandini, "Melting the iceberg: contrast invariance in visual cortex," *Neuron*, vol. 54, no. 1, pp. 11–13, 2007.

[58] H. Xu, H.-Y. Jeong, R. Tremblay, and B. Rudy, "Neocortical Somatostatin-Expressing GABAergic Interneurons Disinhibit the Thalamorecipient Layer 4," *Neuron*, vol. 77, no. 1, pp. 155–167, 2013.

[59] M. W. Spratling, "Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function," *Neural Computation*, vol. 24, pp. 60–103, Dec. 2011.

[60] J. Shi, J. Wielaard, and P. Sajda, "Analysis of a Gain Control Model of V1: Is the Goal Redundancy Reduction?," in *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4991–4994, 2008.

[61] M. Spratling, "A single functional model accounts for the distinct properties of suppression in cortical area V1," *Vision Research*, vol. 51, no. 6, pp. 563–576, 2011.

[62] M. Spratling, "Predictive coding accounts for v1 response properties recorded using reverse correlation," *Biological Cybernetics*, vol. 106, pp. 37–49, 2012.

[63] M. Spratling, "Predictive coding as a model of the v1 saliency map hypothesis," *Neural Networks*, vol. 26, no. 0, pp. 7 – 28, 2012.

[64] R. Rao and D. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, pp. 79–87, 1999.

[65] O. Schwartz and E. Simoncelli, "Natural signal statistics and sensory gain control," *Nature neuroscience*, vol. 4, no. 8, pp. 819–825, 2001.

[66] R. Coen-Cagli, P. Dayan, and O. Schwartz, "Cortical surround interactions and perceptual salience via natural scene statistics," *PLoS Comput Biol*, vol. 8, p. e1002405, 03 2012.

[67] Y. Karklin and M. Lewicki, "Emergence of complex cell properties by learning to generalize in natural scenes," *Nature*, vol. 457, no. 7225, pp. 83–86, 2008.

[68] T. Lochmann, U. A. Ernst, and S. Denève, "Perceptual inference predicts contextual modulations of sensory responses," *The Journal of Neuroscience*, vol. 32, no. 12, pp. 4179–4195, 2012.

[69] M. Zhu and C. J. Rozell, "Sparse coding models demonstrate some non-classical receptive field effects," *BMC Neuroscience*, vol. 11, no. Suppl 1, p. O21, 2010.

[70] A. Koulakov and D. Rinberg, "Sparse incomplete representations: A potential role of olfactory granule cells," *Neuron*, vol. 72, no. 1, pp. 124 – 136, 2011.

[71] P. Berkes, B. White, and J. Fiser, "No evidence for active sparsification in the visual cortex," *Advances in Neural Information Processing Systems*, vol. 22, pp. 108–116, 2009.

[72] V. A. F. Lamme, K. Zipser, and H. Spekreijse, "Figure-ground activity in primary visual cortex is suppressed by anesthesia," *Proceedings of the National Academy of Sciences*, vol. 95, no. 6, pp. 3263–3268, 1998.

[73] I. Finn, N. Priebe, and D. Ferster, "The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex," *Neuron*, vol. 54, no. 1, pp. 137–152, 2007.

[74] A. Angelucci and P. Bressloff, "Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons," *Progress in Brain Research*, pp. 93–120, 2006.

[75] M. Carandini, D. Heeger, and W. Senn, "A synaptic explanation of suppression in visual cortex," *Journal of Neuroscience*, vol. 22, no. 22, pp. 10053–10065, 2002.

[76] M. Zhu, B. A. Olshausen, and C. J. Rozell, "Biophysically accurate inhibitory interneuron properties in a sparse coding network," in *Computational and Systems Neuroscience (Cosyne) Meeting*, 2012.

[77] D. D. Bock, W.-C. A. Lee, A. M. Kerlin, M. L. Andermann, G. Hood, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim, and R. C. Reid, "Network anatomy and in vivo physiology of visual cortical neurons," *Nature*, vol. 471, pp. 177–182, Mar. 2011.

[78] E. Fino and R. Yuste, "Dense inhibitory connectivity in neocortex," *Neuron*, vol. 69, no. 6, pp. 1188 –1203, 2011.

[79] V. Egger, D. Feldmeyer, and B. Sakmann, "Coincidence detection and changes of synaptic efficacy in spiny stellate neurons in rat barrel cortex," *Nature Neuroscience*, vol. 2, pp. 1098–1105, Dec. 1999.

[80] H. Barlow and P. Földiák, "Adaptation and decorrelation in the cortex," in *The computing neuron* (R. Durbin, C. Miall, and G. Mitchison, eds.), pp. 54–72, Addison-Wesley Longman Publishing Co., Inc., 1989.

[81] M. P. Sceniak, M. J. Hawken, and R. Shapley, "Visual spatial characterization of macaque v1 neurons," *Journal of Neurophysiology*, vol. 85, no. 5, pp. 1873–1887, 2001.

[82] Y.-J. Liu, M. Hashemi-Nezhad, and D. C. Lyon, "Dynamics of extraclassical surround modulation in three types of V1 neurons," *Journal of Neurophysiology*, vol. 105, no. 3, pp. 1306–1317, 2011.

[83] T. Naito, O. Sadakane, M. Okamoto, and H. Sato, "Orientation tuning of surround suppression in lateral geniculate nucleus and primary visual cortex of cat," *Neuroscience*, vol. 149, no. 4, pp. 962–975, 2007.

[84] J. Cavanaugh, W. Bair, and J. Movshon, "Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons," *Journal of Neurophysiology*, vol. 88, no. 5, p. 2530, 2002.

[85] V. A. F. Lamme, V. Rodriguez-Rodriguez, and H. Spekreijse, "Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey," *Cerebral Cortex*, vol. 9, no. 4, pp. 406–413, 1999.

[86] D. Stettler, A. Das, J. Bennett, and C. Gilbert, "Lateral connectivity and contextual interactions in macaque primary visual cortex," *Neuron*, vol. 36, no. 4, pp. 739–750, 2002.

[87] P. Garrigues and B. Olshausen, "Learning horizontal connections in a sparse coding model of natural images," *Advances in Neural Information Processing Systems*, vol. 20, pp. 505–512, 2008.

[88] D. Albrecht and D. Hamilton, "Striate cortex of monkey and cat: contrast response function," *Journal of Neurophysiology*, vol. 48, no. 1, pp. 217–237, 1982.

[89] M. Carandini and D. Ferster, "Membrane potential and firing rate in cat primary visual cortex," *Journal of Neuroscience*, vol. 20, no. 1, pp. 470–484, 2000.

[90] P. Dayan and L. Abbott, *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. MIT Press, 2001.

[91] M. Zhu and C. J. Rozell, "Biologically realistic excitatory and inhibitory cell properties emerge from a sparse coding network," *BMC Neuroscience*, vol. 13, no. Suppl 1, p. P55, 2012.

[92] J. Isaacson and M. Scanziani, "How inhibition shapes cortical activity," *Neuron*, vol. 72, no. 2, pp. 231–243, 2011.

[93] S.-H. Lee, A. C. Kwan, S. Zhang, V. Phoumthipphavong, J. G. Flannery, S. C. Masmanidis, H. Taniguchi, Z. J. Huang, F. Zhang, E. S. Boyden, K. Deisseroth, and Y. Dan, "Activation of specific interneurons improves v1 feature selectivity and visual perception," *Nature*, vol. 488, pp. 379–383, Aug. 2012.

[94] N. R. Wilson, C. A. Runyan, F. L. Wang, and M. Sur, "Division and subtraction by distinct cortical inhibitory networks in vivo," *Nature*, vol. 488, pp. 343–348, Aug. 2012.

[95] P. D. King, J. Zylberberg, and M. R. DeWeese, "Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1," *The Journal of Neuroscience*, vol. 33, no. 13, pp. 5475–5485, 2013.

[96] M. J. Rasch, K. Schuch, N. K. Logothetis, and W. Maass, "Statistical comparison of spike responses to natural stimuli in monkey area v1 with simulated responses of a detailed laminar network model for a patch of v1," *Journal of Neurophysiology*, vol. 105, no. 2, pp. 757–778, 2011.

[97] P. Strata and R. Harvey, "Dale's principle," *Brain Research Bulletin*, vol. 50, no. 5¢6, pp. 349 – 350, 1999.

[98] H. S. Meyer, D. Schwarz, V. C. Wimmer, A. C. Schmitt, J. N. D. Kerr, B. Sakmann, and M. Helmstaedter, "Inhibitory interneurons in a cortical column form hot zones of inhibition in layers 2 and 5a," *Proceedings of the National Academy of Sciences*, vol. 108, no. 40, pp. 16807–16812, 2011.

[99] H. S. Meyer, R. Egger, J. M. Guest, R. Foerster, S. Reissl, and M. Oberlaender, "Cellular organization of cortical barrel columns is whisker-specific," *Proceedings of the National Academy of Sciences*, vol. 110, no. 47, pp. 19113–19118, 2013.

[100] H. Markram, M. Toledo-Rodriguez, Y. Wang, A. Gupta, G. Silberberg, and C. Wu, "Interneurons of the neocortical inhibitory system," *Nat Rev Neurosci*, vol. 5, pp. 793–807, Oct. 2004.

[101] W.-C. A. Lee and R. C. Reid, "Specificity and randomness: structure-function relationships in neural circuits," *Current Opinion in Neurobiology*, vol. 21, no. 5, pp. 801–807, 2011.

[102] R. Rao, B. Olshausen, and M. Lewicki, *Probabilistic models of the brain: Perception and neural function*. The MIT Press, 2002.

[103] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[104] A. Charles, H. L. Yap, and C. J. Rozell, "Short term memory capacity in networks via the restricted isometry property," *Neural Computation*, vol. 26, p. 11981235, 2014.

[105] S. Shapero, C. J. Rozell, and P. Hasler, "Configurable hardware integrate and fire neurons for sparse approximation," *Neural Networks*, vol. 45, no. 0, pp. 134–143, 2013.

[106] T. Hu, A. Genkin, and D. B. Chklovskii, "A network of spiking neurons for computing sparse representations in an energy-efficient way," *Neural Computation*, vol. 24, pp. 2852–2872, Aug. 2012.

[107] A. Balavoine, C. Rozell, and J. Romberg, "Convergence of a neural network for sparse approximation using the nonsmooth łojasiewicz inequality," in *Int. Joint Conf. Neural Netw. (IJCNN)*, 2013.

[108] S. Shapero, M. Zhu, J. Hasler, and C. Rozell, "Optimal sparse approximation with integrate and fire neurons," *International Journal of Neural Systems*, vol. 24, no. 05, p. 1440001, 2014.

[109] D. Ringach, "Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex," *Journal of Neurophysiology*, vol. 88, no. 1, p. 455, 2002.

[110] D. V. Buonomano and W. Maass, "State-dependent computations: spatiotemporal processing in cortical networks," *Nat Rev Neurosci*, vol. 10, pp. 113–125, Feb. 2009.

[111] S. B. Laughlin and T. J. Sejnowski, "Communication in neuronal networks," *Science*, vol. 301, no. 5641, pp. 1870–1874, 2003.

[112] A. B. Lee, K. S. Pedersen, and D. Mumford, "The nonlinear statistics of high-contrast patches in natural images," *International Journal of Computer Vision*, vol. 54, no. 1-3, pp. 83–103, 2003.

[113] A. Srivastava, A. Lee, E. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, pp. 17–33, 2003. 10.1023/A:1021889010444.

[114] M. Galarreta, F. Erdlyi, G. Szab, and S. Hestrin, "Cannabinoid sensitivity and synaptic properties of 2 GABAergic networks in the neocortex," *Cerebral Cortex*, vol. 18, no. 10, pp. 2296–2305, 2008.

[115] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM – Association for Computing Machinery*, vol. 58, no. 3, 2011.

[116] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *ArXiv e-prints*, p. doi: 10.1016/j.jsb.2012.10.010, Sept. 2010.

[117] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[118] A. Charles, A. Ahmed, A. Joshi, S. Conover, C. Turnes, and M. Davenport, "Cleaning up toxic waste: removing nefarious contributions to recommendation systems," in *ICASSP 2013*, 2013.

[119] M. Boerlin, C. K. Machens, and S. Denève, "Predictive coding of dynamical variables in balanced spiking networks," *PLoS computational biology*, vol. 9, no. 11, p. e1003258, 2013.

[120] S. B. Hofer, H. Ko, B. Pichler, J. Vogelstein, H. Ros, H. Zeng, E. Lein, N. A. Lesica, and T. D. Mrsic-Flogel, "Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex," *Nat Neurosci*, vol. 14, pp. 1045–1052, Aug. 2011.

[121] A. M. Packer and R. Yuste, "Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: A canonical microcircuit for inhibition?," *The Journal of Neuroscience*, vol. 31, no. 37, pp. 13260–13271, 2011.

[122] W.-p. Ma, B.-h. Liu, Y.-t. Li, Z. J. Huang, L. I. Zhang, and H. W. Tao, "Visual representations by cortical somatostatin inhibitory neuronsselective but with weak and delayed responses," *The Journal of Neuroscience*, vol. 30, no. 43, pp. 14371–14379, 2010.

[123] L. G. Nowak, M. V. Sanchez-Vives, and D. A. McCormick, "Lack of orientation and direction selectivity in a subgroup of fast-spiking inhibitory interneurons: Cellular and synaptic mechanisms and comparison with other electrophysiological cell types," *Cerebral Cortex*, vol. 18, no. 5, pp. 1058–1078, 2008.

[124] H. Adesnik, W. Bruns, H. Taniguchi, Z. J. Huang, and M. Scanziani, "A neural circuit for spatial summation in visual cortex," *Nature*, vol. 490, pp. 226–231, Oct. 2012.

[125] S. Song, P. J. Sjstrm, M. Reigl, S. Nelson, and D. B. Chklovskii, "Highly nonrandom features of synaptic connectivity in local cortical circuits," *PLoS Biol*, vol. 3, p. e68, 03 2005.

[126] K. Ikeda and J. M. Bekkers, "Autapses," *Current Biology*, vol. 16, no. 9, pp. R308 – R308, 2006.

[127] M. Larkum, "A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex," *Trends in Neurosciences*, vol. 36, no. 3, pp. 141–151, 2013.

[128] C. Sámano, F. Cifuentes, and M. A. Morales, "Neurotransmitter segregation: Functional and plastic implications," *Progress in Neurobiology*, vol. 97, no. 3, pp. 277–287, 2012.

[129] C. A. Runyan, J. Schummers, A. V. Wart, S. J. Kuhlman, N. R. Wilson, Z. J. Huang, and M. Sur, "Response features of parvalbumin-expressing interneurons suggest precise roles for subtypes of inhibition in visual cortex," *Neuron*, vol. 67, no. 5, pp. 847–857, 2010.

[130] A. D. Huberman and C. M. Niell, "What can mice tell us about how vision works?," *Trends in Neurosciences*, vol. 34, no. 9, pp. 464–473, 2011.

[131] C. K. Pfeffer, M. Xue, M. He, Z. J. Huang, and M. Scanziani, "Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons," *Nat Neurosci*, vol. 16, pp. 1068–1076, Aug. 2013.

[132] G. Silberberg, "Polysynaptic subcircuits in the neocortex: spatial and temporal diversity," *Current Opinion in Neurobiology*, vol. 18, no. 3, pp. 332–337, 2008. Signalling mechanisms.

[133] C. DiMattina and K. Zhang, "How to modify a neural network gradually without changing its input-output functionality," *Neural Computation*, vol. 22, pp. 1–47, Oct. 2009.

[134] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the theory of neural computation*, vol. 1. Westview press, 1991.

[135] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[136] M. Carandini, J. Demb, V. Mante, D. Tolhurst, Y. Dan, B. Olshausen, J. Gallant, and N. Rust, "Do we know what the early visual system does?," *Journal of Neuroscience*, vol. 25, no. 46, p. 10577, 2005.

[137] A. L. Barth and J. F. Poulet, "Experimental evidence for sparse firing in the neocortex," *Trends in Neurosciences*, vol. 35, no. 6, pp. 345 – 355, 2012.

[138] K. Ohki, S. Chung, Y. Ch'ng, P. Kara, and R. Reid, "Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex," *Nature*, vol. 433, no. 7026, pp. 597–603, 2005.

[139] S. Shoham, D. O'Connor, and R. Segev, "How silent is the brain: is there a "dark matter" problem in neuroscience?," *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, vol. 192, pp. 777–784, 2006. 10.1007/s00359-006-0117-6.

[140] A. S. Charles, *Dynamics and Correlations in Sparse Signal Acquisition*. PhD thesis, Georgia Institute of Technology, 2015.

[141] L. R. Varshney, P. J. Sjstrm, and D. Chklovskii, "Optimal information storage in noisy synapses under resource constraints," *Neuron*, vol. 52, no. 3, pp. 409 – 423, 2006.

[142] F. Gambino, S. Pagès, V. Kehayas, D. Baptista, R. Tatti, A. Carleton, and A. Holtmaat, "Sensory-evoked ltp driven by dendritic plateau potentials in vivo," *Nature*, vol. 515, no. 7525, pp. 116–119, 2014.

[143] E. N. Brown, P. L. Purdon, and C. J. Van Dort, "General anesthesia and altered states of arousal: A systems neuroscience analysis," *Annual Review of Neuroscience*, vol. 34, no. 1, pp. 601–628, 2011. PMID: 21513454.

[144] T. Hu, C. Pehlevan, and D. B. Chklovskii, "A Hebbian/Anti-Hebbian Network for Online Sparse Dictionary Learning Derived from Symmetric Matrix Factorization," *ArXiv e-prints*, Mar. 2015.

[145] Y. Yoshimura, J. L. M. Dantzker, and E. M. Callaway, "Excitatory cortical neurons form fine-scale functional networks," *Nature*, vol. 433, pp. 868–873, Feb. 2005.

[146] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel, "Functional specificity of local synaptic connections in neocortical networks," *Nature*, vol. 473, no. 7345, pp. 87–91, 2011.

[147] L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel, "Functional organization of excitatory synaptic strength in primary visual cortex," *Nature*, vol. advance online publication, pp. –, Feb. 2015.

[148] K. Ohki and R. Reid, "Specificity and randomness in the visual cortex," *Current opinion in neurobiology*, vol. 17, no. 4, pp. 401–407, 2007.

[149] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for machine learning*. Mit Press, 2012.

[150] A. Hyvärinen and P. Hoyer, "Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces," *Neural computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[151] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1106–1114, 2012.

[152] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3361–3368, IEEE, 2011.

[153] J. Mairal, R. Jenatton, F. R. Bach, and G. R. Obozinski, "Network flow algorithms for structured sparsity," in *Advances in Neural Information Processing Systems*, pp. 1558–1566, 2010.

[154] A. Hyvärinen and P. Hoyer, "A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images," *Vision Research*, vol. 41, no. 18, pp. 2413–2423, 2001.

[155] P. Garrigues and B. Olshausen, "Group sparse coding with a laplacian scale mixture prior," *Advances in Neural Information Processing Systems*, vol. 24, 2010.

[156] J. Huang and T. Zhang, "The benefit of group sparsity," *Ann. Statist.*, vol. 38, pp. 1978–2004, 08 2010.

[157] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, eds.), pp. 82–89, Curran Associates, Inc., 2009.

[158] A. Khosrowshahi, *The laminar organization of V1 neural activity in response to dynamic natural scenes*. PhD thesis, University of California, Berkeley, 2011.

[159] U. Köster, J. Sohl-Dickstein, C. M. Gray, and B. A. Olshausen, "Modeling higher-order correlations within cortical microcolumns," *PLoS computational biology*, vol. 10, no. 7, p. e1003684, 2014.

[160] M. R. Cohen and A. Kohn, "Measuring and interpreting neuronal correlations," *Nat Neurosci*, vol. 14, pp. 811–819, July 2011.

[161] L. M. Martinez, Q. Wang, R. C. Reid, C. Pillai, J.-M. Alonso, F. T. Sommer, and J. A. Hirsch, "Receptive field structure varies with layer in the primary visual cortex," *Nat Neurosci*, vol. 8, pp. 372–379, Mar. 2005.

[162] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate it cortex for core visual object recognition," *PLoS computational biology*, vol. 10, no. 12, p. e1003963, 2014.

[163] J. Ichida, L. Schwabe, P. Bressloff, and A. Angelucci, "Response facilitation from the" suppressive" receptive field surround of macaque V1 neurons," *Journal of Neurophysiology*, vol. 98, no. 4, pp. 2168–2181, 2007.

[164] L. Schwabe, K. Obermayer, A. Angelucci, and P. Bressloff, "The role of feedback in shaping the extra-classical receptive field of cortical neurons: a recurrent network model," *Journal of Neuroscience*, vol. 26, no. 36, p. 9117, 2006.

[165] Y. Ahmadian, D. B. Rubin, and K. D. Miller, "Analysis of the stabilized supralinear network," *ArXiv e-prints*, Feb. 2012.

[166] A. Hyvärinen, J. Hurri, and P. Hoyer, *Natural Image Statistics: A probabilistic approach to early computational vision*. Springer-Verlag New York Inc, 2009.

[167] D. Ferster and K. Miller, "Neural mechanisms of orientation selectivity in the visual cortex," *Annual Review of Neuroscience*, vol. 23, no. 1, pp. 441–471, 2000.