STATISTICS OF SINGLE UNIT RESPONSES IN THE HUMAN MEDIAL TEMPORAL LOBE: A SPARSE AND OVERDISPERSED CODE

A Dissertation in Physics
by
Andrew Magyar

© 2015 Andrew Magyar

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

December 2015
The dissertation of Andrew Magyar was reviewed and approved* by the following:

John C. Collins  
Distinguished Professor of Physics  
Dissertation Advisor, Chair of Committee

Dezhe Jin  
Associate Professor of Physics

Lyle Long  
Distinguished Professor of Aerospace Engineering and Mathematics

Patrick Drew  
Assistant Professor of Engineering Science & Mechanics and Neurosurgery

Nitin Samarth  
Professor of Physics  
George A. and Margaret M. Downsbrough Department Head

*Signatures are on file in the Graduate School.
Abstract

The recent discovery of cells that respond to purely conceptual features of the environment (particular people, landmarks, objects, etc) in the human medial temporal lobe (MTL), has raised many questions about the nature of the neural code in humans. The goal of this dissertation is to develop a novel statistical method based upon maximum likelihood regression which will then be applied to these experiments in order to produce a quantitative description of the coding properties of the human MTL. In general, the method is applicable to any experiments in which a sequence of stimuli are presented to an organism while the binary responses of a large number of cells are recorded in parallel. The central concept underlying the approach is the total probability that a neuron responds to a random stimulus, called the neuronal sparsity. The model then estimates the distribution of response probabilities across the population of cells. Applying the method to single-unit recordings from the human medial temporal lobe, estimates of the sparsity distributions are acquired in four regions: the hippocampus, the entorhinal cortex, the amygdala, and the parahippocampal cortex. The resulting distributions are found to be sparse (large fraction of cells with a low response probability) and highly non-uniform, with a large proportion of ultra-sparse neurons that possess a very low response probability, and a smaller population of cells which respond much more frequently. Ramifications of the results are discussed in relation to the sparse coding hypothesis, and comparisons are made between the statistics of the human medial temporal lobe cells and place cells observed in the rodent hippocampus.
# Table of Contents

List of Figures vii

List of Tables viii

Acknowledgments ix

## Chapter 1
Sparse Coding

1.1 Introduction: Sparse coding as an end goal of sensory processing 1
1.1.1 Theoretical advantages of sparse coding 3
1.1.2 Experimental evidence of sparse coding 4

1.2 Concept cells in humans: the Jennifer Aniston neuron 6
1.2.1 Experimental Paradigm 7
1.2.2 Comparison with place cells 8

1.3 Model preliminaries: Measuring sparseness 9
1.3.1 Neuronal Sparsity 10
1.3.2 Binary measures 11
1.3.3 Firing rate measures 11
1.3.4 Experimental Limitations 14
1.3.4.1 Multiple-Neuron Units: Limitations of Spike Sorting 15
1.3.4.2 Undetected Silent Cells 15
1.3.5 Outlook and Goals of the model 15

## Chapter 2
Model Definition and Likelihood Derivation 17

2.1 Introduction 17
2.2 Definition of Model 18
2.2.1 Assumptions 18
2.2.2 Inferential Statistics 19
2.2.3 Single-Unit Response Probability 20
## 2.3 Models Tested

- 2.3.1 One Population ........................................... 22
- 2.3.2 Intermediate Model ....................................... 23
- 2.3.3 Two Populations ........................................... 23
- 2.3.4 Beta Model .................................................. 24

## 2.4 Method of Maximum Likelihood

- 2.4.1 The likelihood function ................................... 25
- 2.4.2 Poisson Approximation .................................... 26
- 2.4.3 Sufficient Statistics ....................................... 27

## 2.5 Chi-squared Analysis ........................................ 29

## 2.6 Multiple Neuron Units ...................................... 30

- 2.6.1 Effective Sparsity .......................................... 31
- 2.6.2 Single-Unit Response ...................................... 31
- 2.6.3 Multi-Unit Predictions ..................................... 32
  - 2.6.3.1 One Active Population .................................. 32
  - 2.6.3.2 Intermediate Model .................................... 33
  - 2.6.3.3 Two Active Populations ................................. 35
  - 2.6.3.4 Beta Distribution ....................................... 36

## Chapter 3

### Model Fits

- 3.1 Introduction .................................................. 39
- 3.2 Failure of One-Population and Intermediate Models ......... 40
  - 3.2.1 One-Population fits ...................................... 40
  - 3.2.2 Intermediate Model Fits .................................. 43
- 3.3 Two-Population Fits .......................................... 43
- 3.4 Beta Fits ...................................................... 45
- 3.5 Multi-Unit Results ........................................... 46
  - 3.5.1 Multi-unit fits: two population model .................... 46
  - 3.5.2 Beta Multi-unit Fits ....................................... 48
  - 3.5.3 Sparsening Effect .......................................... 48
  - 3.5.4 Multi-units did not salvage one-population model ...... 48
- 3.6 Comparison .................................................... 49

## Chapter 4

### Coding Implications

- 4.1 How Sparse is Sparse? ....................................... 56
  - 4.1.1 Small sparsity values ..................................... 57
  - 4.1.2 How close are MTL neurons to GM cells? ................ 57
  - 4.1.3 Silent Cells ................................................. 61
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2 Multiple Models Fit Identical Data</td>
<td>62</td>
</tr>
<tr>
<td>4.2.1 two populations</td>
<td>62</td>
</tr>
<tr>
<td>4.2.2 Beta model vs Two-population model</td>
<td>63</td>
</tr>
<tr>
<td>4.2.3 Other possible models</td>
<td>65</td>
</tr>
<tr>
<td>4.2.4 Speculation on the decoding advantages of a near-power law sparsity distribution</td>
<td>65</td>
</tr>
<tr>
<td>4.3 Beta Distributions arising from preferential attachment processes</td>
<td>66</td>
</tr>
<tr>
<td>4.3.1 Rich get Richer schemes</td>
<td>68</td>
</tr>
<tr>
<td>4.3.2 Bi-layered network model of MTL coding</td>
<td>68</td>
</tr>
<tr>
<td>4.3.3 Growing the network</td>
<td>69</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>71</td>
</tr>
<tr>
<td>Conclusions</td>
<td></td>
</tr>
<tr>
<td>Appendix A</td>
<td></td>
</tr>
<tr>
<td>Derivation of Poisson approximation to likelihood function</td>
<td>73</td>
</tr>
<tr>
<td>A.1 Introduction</td>
<td>73</td>
</tr>
<tr>
<td>Appendix B</td>
<td></td>
</tr>
<tr>
<td>Gamma-Poisson distribution as a limiting case of the beta-binomial distribution</td>
<td>76</td>
</tr>
<tr>
<td>Appendix C</td>
<td></td>
</tr>
<tr>
<td>Beta Distribution as the limiting distribution of preferential attachment</td>
<td>79</td>
</tr>
<tr>
<td>C.1 The node degree distribution</td>
<td>79</td>
</tr>
<tr>
<td>C.2 Deriving the beta distribution</td>
<td>82</td>
</tr>
<tr>
<td>Appendix D</td>
<td></td>
</tr>
<tr>
<td>Error Analysis</td>
<td>84</td>
</tr>
<tr>
<td>Bibliography</td>
<td>86</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Multi-unit response probability for one-population model .......... 34
2.2 Multiple-neuron unit response probabilities for the two-population model ................................................. 36
2.3 Multiple-neuron unit response probabilities for the beta model . . 38

3.1 Hippocampus fits for the four models ................................. 41
3.2 Entorhinal cortex fits .................................................... 50
3.3 Amygdala fits ............................................................. 51
3.4 Parahippocampal cortex fits .......................................... 52
3.5 Histograms of two-population model fits split by population ..... 53
3.6 Beta model sparsity distributions ..................................... 54
3.7 Two-population, multi-unit fits in hippocampus .................. 54
3.8 Beta model, multi-unit fits in hippocampus ....................... 55

4.1 Comparison of predictions between two-population and beta models 62
4.2 Beta-Binomial fits vs binomial fits ................................. 64
4.3 Comparison of $k = 1$ bins between two-population model and the beta model ............................................. 66
List of Tables

1.1 Experimental Sessions ............................................. 7
3.1 Mormann et al. data ................................................. 40
3.2 Best fit parameter values ........................................... 42
3.3 Parameter values for multi-neuron units fits ...................... 46
3.4 Parameters of beta model assuming the presence of multi-units ... 48
D.1 Correlations between two-population model parameters .......... 85
Acknowledgments

I would like to thank John for being a supportive, insightful, and patient advisor throughout this long process, and for setting me to work on such a fascinating problem. It was a pleasure to work with him, and I have grown tremendously under his guidance over these past several years.

I am also extremely grateful for all the people closest to me who have been supportive in various ways—as friends, peers, and mentors. Alex, in particular, through his constant encouragement and quality advice, has kept me strong and on-track.

I especially thank my parents, Mary Jo and Eric, for enkindling my curiosity ever since I can remember, and for providing an endless source of love and support. I owe them everything and more.
Chapter 1  
Sparse Coding

The discovery of concept cells in the human medial temporal lobe, reported in Quian Quiroga et al. [1]—cells that respond reliably and robustly to specific people, landmarks, or objects—has served as the central focus of my research over the past several years. These experiments reveal that the neural code in humans is strikingly sparse, with cells responding to extremely specific stimuli, and that the activity of single neurons can carry tremendous amounts of information about the surrounding environment.

The goal is to develop a statistical method for analyzing this data that will address certain quantitative aspects about the coding properties of the cells in the human MTL. The overarching questions that motivate this research are: How sparse is the neural code in the MTL? and Are the neural response statistics homogeneous or are they widely varying from cell to cell?. In this Chapter, we begin by introducing the idea of sparse coding and summarizing key results that place the concept cell experiments in historical context. Then, the neuronal sparsity is defined and put forward as the appropriate metric for quantifying the MTL code. Finally, this chapter concludes with a discussion of experimental limitations that must be taken into account as the model is developed in the following chapters.

1.1 Introduction: Sparse coding as an end goal of sensory processing

The sparse coding hypothesis states that neural processing of sensory information is organized to produce representations of salient aspects of the environment (people,
objects, landmarks, etc.) using only a small number of strongly activated neurons.

The idea of the sparse coding hypothesis grew out of the classic series of experiments by Hubel and Weisel in the late 1950s through the early 1960s which found cells in the visual cortex of the cat that possess increasingly complex and specific receptive fields (dots to lines to corners to shapes etc.) along the sequential stages of the processing path [2–4]. Extrapolating this discovery along the entire visual pathway, J. Konorski in 1967 [5] and H. Barlow in 1972 [6] independently proposed that cells at the most advanced stages of processing have receptive fields corresponding to behaviorally relevant features of visual scenes, such as individuals, landmarks, particular objects, etc. Konorski referred to these as “gnostic cells” while Barlow called them “cardinal cells.” Both authors speculated that these cells, should they exist, would form the fundamental components of perception.

This line of inquiry established the feed-forward paradigm of sensory processing (see e.g. Chapter 21 in Kandel, Schwartz et al. [7]), in which information flows from the sensory receptors through the CNS in a sequence of stages. At each stage, the sensory information is transformed from the previous stage, with the neurons responding to progressively more complex features of the environment. This means that at each stage, the neurons individually will respond less and less frequently as their receptive fields become more and more specific.

This also marks the genesis of the grandmother cell hypothesis [8], which refers to a hypothetical cell (or group of cells) at the tail end of the visual processing hierarchy in the human brain that would respond only to scenes containing a specific person—e.g. your grandmother. Any scenes or stimuli containing your grandmother would trigger activity in this cell, while all stimuli that do not contain your grandmother would fail to evoke above-baseline activity in the cell. Since its inception, the grandmother cell hypothesis has been extensively criticized, with most researchers favoring a more distributed representation of high-level stimulus features in which the a pattern of activity across multiple neurons would code for specific individuals [9].

Both the controversy over grandmother cells, and the open questions concerning the relationship between single-neuron activity and perception, set in motion a large effort to better understand the highest stages of visual processing. What are the most complex receptive fields in the brain? If not grandmother cells or gnostic cells, then what? In other words, what is the end result of sensory processing?
Addressing these questions with regards to vision requires looking at single neuron recordings from the final stages of the ventral stream. In mammals, this is the inferior temporal lobe (IT), which is the highest purely visual processing area [10,11]. From there, the IT sends massive projections to the medial temporal lobe (MTL) areas [12–14]. The MTL consists of several structures responsible for retrieval and encoding of declarative memory, including the hippocampus [15,16]. Thus, the neurons in the IT and MTL, which receives the processed output of the IT, should have the most complex receptive fields. If grandmother cells exist, they would likely be found in these regions.

Over the last several decades, many studies have found cells with extremely specific response characteristics in a variety of organisms, including humans. Also, the activity in the higher stages has been found to be very sparse. In the following discussion, I will summarize a small part of the converging evidence in favor of the sparse coding hypothesis, which is often described as a compromise between grandmother or local coding and fully distributed coding [17–20]. In a sparse code, the cells respond to a very small proportion of stimuli and have very specific receptive fields, though the fields may not correspond to unique, individual people or objects. For example, a cell in a sparse code may respond to a small set of people, or a small collection of objects.

\[ \text{1.1.1 Theoretical advantages of sparse coding} \]

A code that exhibits population sparseness possesses many favorable properties. Among them include a moderately high representational capacity, higher than local or grandmother codes, but lower than fully distributed codes [17]. For simplicity, let's consider a set of \( N \) binary “on or off” cells responsible for representing features of the environment. This population of cells then has \( 2^N \) distinct activation patterns. A fully distributed code on these \( N \) binary cells is defined as a coding scheme that can in principle utilize all \( 2^N \) patterns of activity to code features—an enormous number for any realistic value of \( N \). On the other hand, a system of \( N \) grandmother cells, in which each feature is coded in the activity of a single cell, is only capable

\[ \text{This property of a cell is referred to as \textit{lifetime sparseness}. A related but distinct definition of sparse coding involves what is called \textit{population sparseness} in which a particular stimulus evokes activity in a small proportion of cells [21,22]. For the remainder of the dissertation, sparse coding is synonymous with lifetime sparseness unless otherwise stated.} \]
of representing $N$ features. In a sparse coding scheme, a small subset of the $N$ cells is active, yielding a high representational capacity due to the combinatorial properties of choosing elements from a large set.

Another advantage of sparse coding is energy efficiency, owing to a small number of simultaneously active neurons [23–25]. In particular, the calculation by Lennie [25] estimates that 1/50 cells at most can be active throughout the brain due to energy constraints at any giving point in time.

Thirdly, sparse codes are easily read compared with distributed codes, though not as easy to read as grandmother codes. In a grandmother code, the activity of a single cell indicates the presence or absence of that concept or feature in the stimulus. Any readout cells interested in a particular feature or set of features would only need connection from the corresponding set of grandmother cells. Distributed codes, on the other hand, in which the presence or absence of a particular feature can only be ascertained by examining the responses of the entire set of cells, requires widespread connections with the readout cells. Sparse codes lie between these two extremes, at a favorable trade off between representational capacity and readability [17].

Returning to Barlow’s vision of hierarchical sensory processing [6], as information is fed along the sensory pathways, redundant or irrelevant information is shed, leaving only features of behavioral significance at the top levels. Furthermore, the top levels of the processing hierarchy (IT, MTL) is where ease of readout is more important than representational capacity. The combination of fewer relevant features, energy constraints, and emphasis on readability give strong reason to suspect sparse coding of behaviorally salient features is the end goal of sensory processing.

### 1.1.2 Experimental evidence of sparse coding

The properties listed above in section 1.1.1 are properties describing a population of cells. In an experimental setting, however, it is difficult to record from a substantial fraction of the available cells in a particular brain region. Thus, many of these theoretical population coding issues have proven difficult to address directly [21].

It is much easier to examine a single cell and to ask for each individual cell “What fraction of stimuli trigger activity in this particular neuron?” rather than to present a single stimulus and ask “How many of the available cells are simultaneously
triggered by this stimulus?” The former is referring to the response properties of individual cells across stimuli, while the latter is referring to the coding properties of a stimulus across cells. If a population exhibits sparse coding (i.e. a small fraction of cells are active for each stimulus), then it follows that at least some (not necessarily all) of the cells within the population must respond to a very small proportion of stimuli. Such neurons are referred to as sparse cells. The detection of such cells is then taken as experimental evidence of sparse coding.

Experimental detection of sparse codes involves identifying cells that are highly selective, responding to a very small proportion of complex stimuli. Examples have been found in a diverse array of species across many sensory modalities. Examples include the odor-specific responses of Kenyon cells in locusts [26], V1 cells in cats [27] and mice [28]. Also, the RA projecting neurons of the HVC in zebra finches are extremely sparse–firing only at a single point during a song [29].

The visual system in non-human primates has been extensively studied over the course of several decades, revealing neurons possessing a variety of complex response characteristics. In 1969, shortly after Konorski proposed the gnostic cell idea, Gross et al. detected IT neurons within macaques that responded to the presence of a face in the visual field [30]. Following this, many other labs also found face-sensitive cells in monkeys at the higher stages of visual processing [31–36]. In addition to faces, studies have found cells that respond to the presence of shapes [37], objects [10,11,38], and hands [31].

Efforts to capture the statistics of the neural responses in the primate visual system began in earnest with the study by Rolls and Tovee in 1995 [34]. Their measure of neural sparseness is the ratio \[ a = \frac{\langle r \rangle^2}{\langle r^2 \rangle} \] where \( r \) is the firing rate of the cell in response to a given stimulus and the average is taken over all stimuli presented during the experiment. This measure ranges from zero to one. For a sparse code, the neuron will respond to few stimuli with a high firing rate. Thus, the average firing rate would be small, but the variance would be high. Therefore, sparse codes are expected to have \( a \) closer to zero, while distributed codes would have \( a \) closer to one. This coding metric has been widely applied to cells in the IT [27,34,39–41], suggesting a sparse, but far from grandmother code. The limitations of the Rolls-Tovee measure are discussed at length in section 1.3.3.

Electrophysiological studies that investigate the neural code in humans are far less common than in non-human primates due to the unique challenges associated
with clinical neurosurgery. The majority of single unit recordings in humans come from electrodes implanted predominately in the MTL of epileptic patients [42].

Early studies reveal similar results as those in monkeys. Single units recordings in humans have found a variety of cells that respond to complex stimuli such as faces [43–45], object categories [46,47], word pairs [48], and emotional scenes [49].

In humans, striking evidence of sparse coding has been observed in the medial temporal lobe in a series of experiments [1,44,46,50,51] including the concept cells reported in Quian Quiroga et al. [1] which were observed to respond to stimuli related to a single concept (e.g. the “Jennifer Aniston neuron”) out of nearly 100 other concepts presented. These cells reported in the human MTL are much more sparse compared with cells in the monkey IT [52,53].

Developing methods to analyze the concept cell experiments is the central goal of this dissertation. To proceed, I will discuss the results in detail.

1.2 Concept cells in humans: the Jennifer Aniston neuron

Perhaps the most striking examples of sparse cells in humans comes from the experiment reported by Quiroga et al. [1] in which neurons were detected in the MTL with explicit, invariant responses to diverse images associated with a particular concept. For example, one cell was found to produce a highly elevated firing rate in response to multiple images of Jennifer Aniston. Another responded exclusively to images of Halle Berry, including the character string “Halle Berry” presented on a computer screen. A later study by the same group found that these invariant, explicit responses often extended across multiple sensory modalities [51]. These neurons, called “concept cells” [54] typically responded only to a single concept, or a small number of concepts out of roughly 100 presented. These cells give strong evidence for the human MTL employing a very sparse code to represent the high-level features of the environment.

This discovery of cells that respond robustly and sparsely to stimuli associated with particular people reinvigorated the debate over grandmother cells [9,55–61]. While the concept cells display individual-specific responses, the experiment by Quiroga et al. alone does not demonstrate that these are grandmother or
Stimuli

<table>
<thead>
<tr>
<th>Units</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>...</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

(a) Screening Session

Table 1.1. Tables showing hypothetical experimental results from [1]. During the screening sessions (a), a large number of stimuli (images of people, objects, landmarks, animals, etc) were presented. Each letter indicates a distinct concept identity. The binary entries indicate whether or not the unit responds to the stimulus. The rightmost column, K is the sum of the rows, i.e. the number of stimuli that trigger each cell. During the testing sessions (b), the concepts which triggered a response in at least one cell e.g. b and d were shown with multiple viewings, labeled by the subscripts. For example, if b is Jennifer Aniston, b₁, b₂ etc correspond to different images of her, or perhaps her written name.

gnostic cells. A grandmother cell would only respond to one unique concept out of all concepts known to the person, whereas only 100 were tested. Just because a cell responds to one out 100 concepts does not necessarily imply that it responds to one concept out of all concepts.

1.2.1 Experimental Paradigm

The results reported in Quian Quiroga et al. [1] come from electrode recordings in several patients taken during two different sessions. The first is the screening session, shown in Table 1.2.1 (a), in which images of familiar celebrities, landmarks, and objects are presented. About 100 pictures were shown to a patient, each containing
a distinct object (human or otherwise) or objects. Each picture was presented 6 times. The threshold for defining a response was set by the experimenters [1], who required the firing rate of a neuron after presentation of a stimulus to be more than a threshold amount (usually 5 standard deviations) above the baseline pre-stimulus firing rate of the same neuron. The firing rate for the stimulus was averaged over its presentations.

For the images that trigger a response in at least one of the units during this session (say, an image of Bill Clinton), multiple different images of that concept are then shown during the following testing session shown in Table 1.2.1(b). This session explored the responses to further stimuli concerning the same object that caused response(s) in the screening session. These might include very different pictures of the same person, or a written or spoken name. It was often found that the response was conceptual, e.g., multi-modal [51].

Since we are concerned with the statistics of the neural responses to random stimuli, we will be analyzing the data recorded during only the screening sessions, when distinct concepts are presented in each stimulus.

1.2.2 Comparison with place cells

The concept cells of the human MTL and the place cells of the rodent hippocampus [62] share many common characteristics, as outlined in the review by Quian Quiroga [54]. For example, both are highly selective, located in analogous brain regions, and exhibit binary type responses. The concepts cells respond with a greatly increased firing rate to one or a small number of concepts out of 100 presented, and the place cells responding with highly elevated firing rates only when the rodent is in a particular location of the environment [63]. Furthermore, even in the absence of direct visual input, activation of both concept cells [45] and place cells [64] can occur during recall of memories.

These and other similarities lead R. Quian Quiroga to propose that concept cells and place cells are two different manifestations of similar coding processes [54]:

Both concept cells and place cells can be linked to memory process, and the difference between them may simply reflect the different types of stimuli that are salient to each species: whereas for humans it is important to recognize faces (among other things) and to associate
Recent studies have revealed that single place cells can pick up multiple place fields when the rat is allowed to navigate a large environment [65–69]. Statistical analysis of the distribution of the number of place fields per cell reveal that the number of place fields are skewed [69]. There are some place cells with multiple place fields while most have only one field, but the number of cells with multiple fields exceeds the number that would be expected if all place cells recruited place fields at the same rate. In other words, some place cells recruit place fields at a high rate while some must recruit at a lower rate. This skewed distribution of place fields per cell is one example of accumulating observations of skewed behavior in a variety of contexts [70,71].

Similarly, concept cells have been found to possess multiple “concept fields” i.e. the concept cells can respond to multiple individuals, landmarks, objects, etc [1,50,51]. Likewise, the distribution of concept fields per cell has been shown to be skewed by Collins and Jin [72]. The methods of that paper will be extended in this dissertation, following the papers by the author and Collins in [73], and in Magyar [74]. It will be shown that the distribution of concept fields per cell and place fields per cell are strikingly similar, lending further credence to Quian Quiroga’s proposition that they arise from similar mechanisms in the MTL.

1.3 Model preliminaries: Measuring sparseness

Inspired by the sparse coding hypothesis, the grandmother cell debate, and the emerging interest in skewed distributions across various neural systems, the two primary questions to be addressed by the model to be built in the following chapter are: How sparse is the human MTL code? and Are the neural responses of the MTL skewed or homogeneous?

As described above, the concept cell experiments cannot directly test the grandmother cell hypothesis. To measure whether a particular cell implements local coding for a concept, one would need presentation of stimuli that cover a large fraction of the concepts known to the subject. This is evidently far beyond current experimental capabilities, at least. Measurements such as those in [1,50] use stimuli corresponding to only about 100 distinct entities (people, famous buildings, etc).
In the novel statistical method presented, the neurons are postulated to possess a distribution of sparsity values, which we define as the fraction of stimuli that elicit an above-threshold response. The neurons and stimuli in the measurements as being a sample. The vast majority of the neurons are found to respond on average to much less than one stimulus in the approximately 100 that are presented. A direct measurement on a single neuron would need thousands of conceptually distinct stimuli to achieve the same result. The effectiveness of the method arises from the large number of neuron-stimulus trials, i.e., the number of stimuli times the number of neurons, which is in the hundreds of thousands for the data of [1,50].

To proceed, I introduce the neuronal sparsity as the measure to be used. I then briefly discuss why binary measures are appropriate for these cells, and why the commonly used Roll-Treves measure [34] is misleading for this kind of data. I then mention two experimental limitations the model must take into consideration: the sorting of cells into single units and multiple-neuron units, and the silent cell issue.

The following sections of this chapter closely follow the paper Magyar and Collins [73], published in 2015, as well as the paper Magyar [74] which as of writing this, is currently in preparation.

1.3.1 Neuronal Sparsity

To model the MTL data presented in Quian Quiroga et. al [1] and Mormann et al [50], the total fraction of stimuli that elicit a response in a particular neuron, that we term the neuronal sparsity, $\alpha$ is used [73]. This measure assumes a binary code in which the firing rate patterns of neurons can be interpreted as on-or-off.

$$\alpha = \frac{\# \text{ stimuli triggering a response}}{\text{total \# of stimuli}}$$ (1.1)

This definition of sparsity is a property of the neuron itself, and is the total probability that the cell responds to a randomly chosen stimulus. It is not necessarily equal to the fraction of stimuli that elicit a response during a particular experiment in which only a small subset of possible stimuli are presented, as used in [75] for example. If a particular cell remains unresponsive during the presentation of 100 randomly selected images, then its sparsity is not necessarily 0, it may instead have a very low but non-zero sparsity. In other words, $\alpha$ is a quantity that must be inferred from a particular experiment rather than directly calculated.
1.3.2 Binary measures

This is appropriate for neurons which have some fairly low baseline level of activity, but have a substantially higher activity under particular rather specific circumstances, with fairly few border-line cases, as for the pyramidal neuron whose responses are shown in Fig. 3A of [75]. Similar cases where binary classification is sensible include HVC_{RA} neurons for zebra finches in [29].

In contrast typical interneurons exhibit a more uniform and continuous distribution of firing rates, e.g., the interneuron shown in other graphs in Fig. 3A of [75]. For such a neuron, there is no stereotypical responsive and unresponsive activity, and so its sparsity will depend strongly upon the exact threshold chosen to define a response.

Whether the spiking rate was high or low was generally consistent across presentations — see, for example, the raster plots in Fig. 1 in [50] and Figs. 1a and 2a in [1] — although not perfectly so. The firing rate for a stimulus found to cause an above-threshold response was typically well above the threshold, while for a stimulus with a below-threshold response the firing rate was typically well below threshold. That and the conceptual nature of the response indicate that the choice of a binary categorization (responsive or not) is sensible and robust for the neurons studied.

1.3.3 Firing rate measures

This section follows work published in Magyar and Collins [73]. Here, it will be shown that the firing rate measure proposed by Rolls and Tovee [34] could be dominated by the baseline firing rate of the MTL neurons measured in [1], and it could be insensitive to the responsive properties of these cells. A similar conclusion was reached in Quiroga et al. [76] using numerical simulations.

An alternative measure of sparseness proposed by Rolls and Tovee [34] and reviewed by Rolls and Treves [77], is calculated from the firing rates of neurons in response to stimuli. The average sparseness reported [77] with this definition was $0.34 \pm 0.13$ (for hippocampal spatial view cells in the macaque hippocampus). This is dramatically different than the value predicted by Eq (1.1), as will be shown in the following two chapters. It is worth understanding how this difference might arise (aside from a conceivable difference between species). It will now be demonstrated.
that the Rolls-Treves definition can be very misleading as to the nature of neural coding.

They define the sparseness of a neuron in response to a sample of $S$ stimuli as

$$ a_{RT}^s = \frac{\langle r \rangle^2}{\langle r^2 \rangle} = \frac{\left( \sum_{j=1}^S r_j / S \right)^2}{\left( \sum_{j=1}^S r_j^2 / S \right) / S^2}, \quad (1.2) $$

where $r_j$ is the mean firing rate of the neuron in response to stimulus $j$ and $\langle \ldots \rangle$ denotes an average of a quantity over all presented stimuli. In the case that the neuron is exactly binary in its responses, i.e., it gives a large response with some fixed firing rate to some stimuli and is exactly silent for all other stimuli, this definition is the unique combination of $\langle r \rangle$ and $\langle r^2 \rangle$ that matches our definition of sparsity in Eq (1.1). However, as we will now show, Rolls-Treves sparseness and our thresholded sparseness can have widely different values if the neurons are not strictly binary.

We first observe that Rolls and Treves only reported a value of their sparseness averaged over all neurons. Now Ison et al. [75] in their Fig. S2 showed the distribution across neurons of various measures of sparseness and selectivity. There is a wide range of Rolls-Treves sparseness from less than 0.1 (the most common) to another peak close to 1 (for putative interneurons). The average of this distribution is quite misleading as to the properties of individual neurons. Furthermore, all the distributions in that figure are measures of sparseness with respect to the presented stimuli, and no attempt is made to infer an underlying sparseness or selectivity defined with respect to a whole class of stimuli, such as we do here and as Waydo et al. did in [78].

It is well known that many (but not all) hippocampal neurons respond strongly to certain specific behaviorally relevant stimuli or situations, while responding weakly or not at all at other times. The data we analyzed simply give a particularly notable example. Let us refer to the situations to which a cell responds strongly as “on-target” and the other situations as “off-target”. Suppose, that on-target responses are very rare, as we will find, while off-target firing is of a lower but sometimes non-zero rate.

We now construct a specific class of models in which a particular choice of threshold can make it quite unambiguous as to the on- and off-target responses, but
in which the Rolls-Treves sparseness is dominated by properties of the off-target firing and is almost insensitive to the on-target responses. The model is intended merely to be a reasonable counterexample, to show how a high average value of Rolls-Treves sparseness can be compatible with a very low sparsity in our sense.

In this model each neuron has a categorical response to a certain class of stimuli, with probability $\alpha$. We postulate a pseudo-binary model in which the off-target and on-target firing each comes from a different distribution of firing rates, $P_0(r)$ and $P_1(r)$ respectively.\(^2\) Then the total distribution is a mixture:

$$P(r) = (1 - \alpha)P_0(r) + \alpha P_1(r).$$ (1.3)

We let $r_0$ and $\Delta r_0$ be the mean and standard deviation of $P_0$, and let $r_1$ and $\Delta r_1$ be the same quantities for $P_1$. We arrange the distributions $P_0$ and $P_1$ so that on-target responses can be detected adequately reliably by applying an appropriate threshold $r_{th}$. We need

$$\int_{r_{th}}^{\infty} P_0(r) dr \ll \alpha,$$ (1.4)

so that the threshold is sufficiently far out on the tail of $P_0$ that there are many fewer many false positives than true positives. But we also must keep $r_{th}$ not too high

$$\int_0^{r_{th}} P_1(r) dr \ll 1.$$ (1.5)

so that on-target stimuli are reliably detected. For arbitrarily small $\alpha$ distributions obeying these conditions can easily be constructed, at a minimum by having them be non-overlapping.

A calculation yields the Rolls-Treves sparseness:

$$\alpha_{RT}^s = \frac{1}{1 - \alpha} \left( 1 + \frac{\alpha}{1 - \alpha} \frac{r_1}{r_0} \right)^2 - \frac{1}{1 - \alpha} \left( 1 + \frac{(\Delta r_0)^2}{r_0^2} \right) + \frac{\alpha}{(1 - \alpha)^2} \left( \frac{r_1^2}{r_0^2} + \frac{(\Delta r_1)^2}{r_0^2} \right).$$ (1.6)

This reproduces the value $\alpha$ in the case of binary neurons, i.e., where the standard deviations are negligible and the limit $r_0 \to 0$ is taken. But if instead we take

\(^2\)Analysis of a large, unbiased set of neurons in the rat MTL reported in [79] suggests a log-normal distribution for both $P_0(r)$ and $P_1(r)$. However, the exact forms of the distributions $P_0(r)$ and $P_1(r)$ are not important in this analysis, so long as they are sufficiently non-overlapping.
the situation that $\alpha$ is very small, while keeping $r_0$ non-zero, then the Rolls-Treves sparseness approaches $\frac{1}{1 + \Delta r_0^2/r_0^2}$. This is just the value calculated purely from the off-target distribution. Hence the Rolls-Treves sparseness can be dominated by its off-target value, while by use of a suitable threshold we have reliable discrimination between on-target and off-target stimuli, with $\alpha$ having the small value given by our definition.

The average sparseness reported [77] with this definition was $0.34 \pm 0.13$ (for hippocampal spatial view cells in the macaque hippocampus). We will show in Chapter 3 that cells in the human hippocampus possess a mean sparsity of $\approx 10^{-3}$, dramatically less what is found in the macaque hippocampus using Eq 1.2. Thus, the choice of sparseness measure is crucial. In this situation, there is no contradiction between a rather high value of Rolls-Treves sparseness, such as 0.3 and extremely low values of our sparsity. Given the behavioral significance of the above-threshold responses in [1], it is sparsity in our sense that appears more relevant to understanding the nature of the corresponding conceptual coding.

### 1.3.4 Experimental Limitations

There are two complications inherent to single-unit recordings that must be taken into consideration as we proceed with the model development. The first is the fallibility of spike-sorting algorithms used to split the spike trains detected by an electrode into spike trains produced by individual neurons. It is often the case that two or more neurons produce spikes that are indistinguishable, and so what appears to be the activity of a single cell actually may represent the accumulated activity of two or more cells. The second issue is the presence of a huge population of completely silent cells that are never detected by the electrode, despite being within range. The silent cell issue implies that the units detected constitute a highly biased sample, consisting of only the most active and responsive units.

Finally, it is worth emphasizing that in the experiment to be analyzed, only a very small sample of the whole neural population is reported. Thus, the exact numerical values reported in this dissertation should be interpreted as reasonable ballpark estimates. More emphasis will be placed on the qualitative features of the distributions studied, rather than upon the exact vaues of the fitted parameters.
1.3.4.1 Multiple-Neuron Units: Limitations of Spike Sorting

In the data that we analyze from [50], units were identified from electrode recordings using spike sorting techniques which cannot always distinguish individual neurons. Thus, the firing patterns reported for some units represent the aggregate firing of multiple neurons rather than for a single neuron. The simplest version of our model ignores this distinction, and assumes that each recorded unit consists of only a single neuron. But, as we will explain in Sec. 2.6, the general principles of our methods apply perfectly well at the unit level. We will show how to transform from a neural-level model to a unit-level model, and we will see how to interpret numerical results of fits at the unit level in terms of single neuron properties. This will result in no change in our qualitative results, but a strengthening of our conclusions about the presence of a large number of very sparsely responding neurons.

1.3.4.2 Undetected Silent Cells

Another fundamental limitation of single unit electrode recordings is the inability to detect totally silent cells which emit no spikes during the recording process [80–82]. The cells that are detected by any particular electrode constitute a small minority of the total cells that are within its range. This “neural dark matter” is suspected to represent a substantial proportion of all neurons that are within the range of an electrode. Thus, any cells that are reported in such studies constitute a biased sample of only the most active minority, and so any calculation of sparsity should be treated as an upper bound on the true sparsity.

1.3.5 Outlook and Goals of the model

Now that we have chosen the neuronal sparsity in Eq (1.1) as the metric for the sparseness of neural responses reported in the concept cell experiments described above, we are prepared to describe the procedure for fitting the data using inferential statistics, discussed in detail in Chapter 2. We seek to estimate the distribution of $\alpha$ across the population of human MTL cells. To proceed, we first need to select a family of distributions $D_{\bar{\theta}}(\alpha)$ where $\bar{\theta}$ are the parameters of the distribution chosen.

Distributions that will be explored are single population models where $D_{\bar{\theta}}(\alpha)$ is given by a delta function (this model tries to fit the data assuming that all cells
have the same $\alpha$), two population models, where the distribution is given by a weighted sum of two delta functions (implying that the neurons are split into two populations, each with a distinct sparsity), and beta models, where $D_\theta(\alpha)$ is given by the beta distribution.

After the various models are chosen, they will be fit to the data presented in Mormann et al. [50] using the maximum likelihood procedure. Then, the $\chi^2$ test for goodness of fit will be introduced to assess the maximum likelihood fits. Finally, the chapter concludes by developing methods for incorporating multiple-neuron units discussed above in section 1.3.4.1.

Then, in Chapter 3, the best fits will be presented, with the corresponding sparsity distributions, $D_\theta(\alpha)$. It will be evident that homogeneous distributions of sparsity cannot replicate the data, and that skewed distributions are required. The models that successfully fit the data i.e. two-population model and the beta model, indicate a large proportion of extremely sparse neurons–with over 90% possessing a sparsity less than $10^{-3}$. The inclusion of multiple-neuron units serves to lower the sparsity even further. The beta distribution fits have approximate power-law behavior at low sparsities.

Finally in Chapter 4, the fits are discussed in light of the ideas presented above. The successful fits and the corresponding sparsity distributions are discussed in relation to grandmother cell coding, decodability, and the conceptual nature of the neural responses. The limitations of the model are described, and the presence of large proportions of silent cells are analyzed. Finally, a preferential attachment mechanism is proposed as a possible origin of the beta distributed sparsity.
Chapter 2
Model Definition and Likelihood Derivation

2.1 Introduction

In this chapter, I will develop the probabilistic models presented in Magyar and Collins [73] and Magyar [74]. The models build off work by Collins and Jin [72], and they are used to predict the outcome of an experiment of the type reported in Quiñon Quiroga et al. [1], in which $N$ independent, binary neurons are recorded in parallel as $S$ stimuli are presented.

The results of the experiment are given by the numbers $n_0, n_1, \ldots, n_S$ where $n_k$ is the number of recorded cells that respond to $k$ out of the $S$ stimuli presented. We assume that neuron $j$ has a sparsity $\alpha_j$ defined for that particular neuron by Eq. (1.1). The goal of the model is to estimate the distribution $D_\bar{\theta}(\alpha)$ with parameters of sparsity values across the population of human MTL neurons based upon the reported $n_k$ data in Mormann et al. [50].

To do this, I proceed along these steps:

1. Define the model and state assumptions (constant sparsity, independent neurons, random stimuli, etc.).

2. Derive the probability that a randomly chosen unit responds to $k$ stimuli, $\epsilon_k(\bar{\theta})$, in terms of the sparsity distribution $D_\bar{\theta}(\alpha)$.

3. Use $\epsilon_k(\bar{\theta})$ to determine the model expectation values for the bin counts for the experimental outcome, $n_k^*(\bar{\theta})$. 
4. Define the different distributions $D_{\bar{\theta}}(\alpha)$ studied in the following chapter.

5. Describe the maximum likelihood procedure for finding $\bar{\theta}_0$, the model parameters that best fit the prediction, $n_k^*(\bar{\theta})$, to the data, $n_k$.

6. Expand the fitting procedure to incorporate multi-unit activity.

7. Compare our procedure to other analyses performed on the same data.

2.2 Definition of Model

The task in this section is to carefully present the assumptions that underlie the model and to derive the single unit response probability $\epsilon_k(\bar{\theta})$, defined as the probability that a randomly selected neuron responds to $k$ out of the $S$ stimuli.

2.2.1 Assumptions

We assume:

- binary neural responses
- The recorded neurons are statistically independent of each other.
- The set of $S$ stimuli presented constitute a sample from some large “universe” of all possible stimuli
- The set of $N$ neurons recorded constitute a sample from the full population of MTL neurons
- The probability that neuron $j$ responds to a random stimulus is equal to its sparsity, $\alpha_j$

In the experiment analyzed, the stimuli consist of visual images each containing a distinct object (human or otherwise) or objects. Each picture was presented 6 times. The threshold for defining a response was set by the experimenters [1], who required the firing rate of a neuron after presentation of a stimulus to be more than a threshold amount (usually 5 standard deviations) above the baseline pre-stimulus firing rate of the same neuron. The firing rate for the stimulus was averaged over its presentations.
Whether the spiking rate was high or low was generally consistent across presentations — see, for example, the raster plots in Fig. 1 in [50] and Figs. 1a and 2a in [1] — although not perfectly so. The firing rate for a stimulus found to cause an above-threshold response was typically well above the threshold, while for a stimulus with a below-threshold response the firing rate was typically well below threshold. That and the conceptual nature of the response indicate that the choice of a binary categorization (responsive or not) is sensible and robust for the neurons studied.

The statistical independence of the recorded neurons was verified experimentally in [1].

2.2.2 Inferential Statistics

The aim of the model is not to characterize the data itself, with its limited sample of neurons and stimulus. Rather, we wish to infer response properties for the “universe” of all neurons in particular brain regions and all relevant stimuli. Our methods also provide tools to correct for the bias in reporting responsive neurons in the presence of many neurons that, within the limits of a particular experiment, are either unresponsive or silent [80,83].

One advantage of using statistical inference to estimate the underlying distribution of sparsity, $D(\alpha)$, is that the form of the distribution can suggest a particular generating mechanism by looking into the possible processes or systems in which similar distributions arise. For example, if a variable is found to follow a Gaussian distribution, then it might suggest the variable represents the sum of a large number of independent random variables.

For the models tested in this dissertation, two of them are consistent with the data: the two population model described in Sec. 2.3.3, and the beta model introduced in Sec. 2.3.4. The two-population model, with cells displaying two distinct response characteristics, immediately suggests two populations of cells in the MTL with some sort of functional or anatomical distinction. For example, Kozhevnikov et al. [84] found considerably different response characteristics in the RA-projecting cells vs the X-projecting cells in the HVC of the zebra finch during a song production. This is discussed in further detail in Sec. 4.2.

Beta distributions, on the other hand, commonly arise via preferential attach-
ment processes and various “rich-get-richer” schemes [85–87]. This suggests the possibility that the neural MTL code might grow in a similar fashion as new concepts are encoded, with new concepts tending to become encoded on cells that have already coded for many previous concepts—an idea that will be explored in 4.3.

2.2.3 Single-Unit Response Probability

If neuron $j$ responds to each of $S$ random stimuli with probability $\alpha_j$, then the probability of $K_j = k_j$ responses is given by the binomial distribution:

$$P(K = k \mid \alpha_j) = \binom{S}{k} \alpha_j^k (1 - \alpha_j)^{S-k}. \tag{2.1}$$

If $\alpha_j$ is sampled from some distribution $D_{\hat{\theta}}(\alpha_j)$, then the total probability that neuron $j$ responds to $K = k$ stimuli, $\epsilon_k(\bar{\theta})$, is given by:

$$P(K = k) = \binom{S}{k} \int_0^1 d\alpha D_{\hat{\theta}}(\alpha) \alpha^k (1 - \alpha)^{S-k} \equiv \epsilon_k(\bar{\theta}). \tag{2.2}$$

Eq. (2.2) is a mixture distribution, where the binomial distribution of Eq. (2.1) is mixed by the sparsity distribution. Choice of $D_{\hat{\theta}}(\alpha)$ and Eq. (2.1) define the model being used.

The expectation values for the bin counts, $n^*_k(\bar{\theta})$ are calculated using Eq. (2.2). The probability that each neuron responds to $k$ stimuli is given by $\epsilon_k(\bar{\theta})$. Thus, the probability that we observe $n_k$ is given by a binomial distribution with expectation $n^*_k \equiv \langle n_k \rangle = N \epsilon_k$ and standard deviation $\sqrt{N \epsilon_k (1 - \epsilon_k)}$. Note the dependency of $\epsilon_k$ and $n^*_k$ on the model parameters $\bar{\theta}$ is dropped for simplicity.

More explicitly, the model prediction is given by

$$n^*_k = N \epsilon_k \pm \sqrt{N \epsilon_k (1 - \epsilon_k)}. \tag{2.3}$$

The standard deviation determines the expected amount of variation in the bin counts during repetitions of the experiment, assuming that during each experiment, the outcomes are sampled from the model.

In the following section, we use the method of maximum likelihood to find the
values of parameters $\bar{\theta}$ that best fit the expected values $n_k^*$ to the observed values $n_k$. In other words, if the observed data were sampled from the model with some probability, then maximum likelihood is used to find the parameters of the model such that the probability of sampling the data is the highest.

### 2.3 Models Tested

1. **Single-population model**: All neurons have the same sparsity, $\alpha_0$. This model has one parameter. In this case,

$$D_{\alpha_0} (\alpha) = \delta (\alpha - \alpha_0). \quad (2.4)$$

2. **Intermediate model**: Two populations present, one active, and one silent. The active population has sparsity $\alpha_D$ and abundance $f_D$, while the silent population has $\alpha_S = 0$, and abundance $1 - f_D$. This model has two parameters, and the sparsity distribution is given by

$$D_{\alpha_D, f_D} (\alpha) = (1 - f_D) \delta (\alpha) + f_D \delta (\alpha - \alpha_D). \quad (2.5)$$

3. **Two-population model**: Two active populations are present, one with parameters $\alpha_D$ and $f_D$, and one ultra-sparse population with parameters $\alpha_{US}$ and $f_{US} = 1 - f_D$. This model has three parameters. The sparsity distribution is given by:

$$D_{\alpha_{US}, \alpha_D, f_D} (\alpha) = (1 - f_D) \delta (\alpha - \alpha_{US}) + f_D \delta (\alpha - \alpha_D). \quad (2.6)$$

4. **Continuous Beta Model**: Neurons are not split into populations. Instead neuronal sparsity is assumed to have a sparsity sampled from a continuous beta distribution. The beta distribution has two parameters, $a > 0$ and $b > 0$, with PDF given by:

$$D_{a, b} (\alpha) = \frac{\alpha^{a-1} (1 - \alpha)^{b-1}}{B(a, b)} \quad (2.7)$$
Above, \( \delta (x) \) is the Dirac-delta function, and \( B (a, b) \) is the beta function defined by

\[
B (a, b) = \int_0^1 dx \, x^{a-1} (1-x)^{b-1}
\]

For each model, we first produce the best fit to the data in [50], by using a maximum likelihood estimate (MLE) for the parameters [88]. We will do this for each of the four regions of the MTL for which data is reported.

We begin with the simplest model, the one-population model, which treats all neurons as having the same sparsity. In Sec. 3.2 the one population model will be shown to produce bad fits in all four regions of the human MTL for which data is reported. To improve the model, we include a second population of cells which remain totally silent \((\alpha = 0)\) in order to account for the heavily populated zero bin. We show that this also fails to predict the data.

We then improve the model further by adding a second active population, so that the MTL neurons are assumed to belong to one of two possible active population. The first active population is labeled US for “ultra-sparse” as the fits will indicate that these neurons fire extremely rarely, and a second population labeled D for “distributed.” The distributed population are neurons that fire more frequently, oftentimes in response to multiple stimuli. This model will be shown in Sec. 3.3 to produce successful fits in two of the four regions: the hippocampus and the entorhinal cortex.

Finally, we try a continuous distribution of sparsity, given by the beta distribution. We show this model to be successful in fitting all four regions in Sec. 3.4

### 2.3.1 One Population

First, we use the model with a single population of neurons with sparsity \( \alpha_0 \). For this model,

\[
D_{\alpha_0} (\alpha) = \delta (\alpha - \alpha_0)
\]

Then Eq. (2.2) becomes a binomial distribution:

\[
\epsilon_k = \binom{S}{k} \alpha_0^k (1 - \alpha_0)^{S-k}.
\]
This is the simplest model, with the single parameter $\tilde{\theta} = \alpha_0$, as it treats all neurons identically. The assumption that all neurons have identical sparsities is made in [78]. In this case, we would expect the data to follow a binomial distribution. It will be shown that this model does not fit the data reported in [50] and that uniform sparsity cannot be assumed in the human MTL.

### 2.3.2 Intermediate Model

A simple and natural improvement to the single population model is to add a second population of completely silent neurons that do not respond to any of the stimuli used — cf. [80]. For the active population, let $\alpha_D$ be its sparsity, and let $f_D$ be its fractional abundance. Then Eq. (2.2) becomes

$$
\epsilon_k = (1 - f_D) \delta(\alpha) + f_D \left( \frac{S^k}{k} \right) \alpha_D^k (1 - \alpha_D)^{S-k},
$$

(2.11)

for $k \geq 1$. This model also will be shown to produce poor fits (though greatly improved over the single population model) to the human MTL data.

### 2.3.3 Two Populations

Therefore our final model uses two active populations. One population we call the distributed population, with a sparsity $\alpha_D$ and fractional abundance $f_D$. The other population we call the ultra-sparse population with sparsity $\alpha_{US}$. The fractional abundance of the ultra-sparse population is $f_{US} = 1 - f_D$. Then Eq. (2.2) becomes

$$
\epsilon_k = (1 - f_D) \left( \frac{S^k}{k} \right) \alpha_{US}^k (1 - \alpha_{US})^{S-k} + f_D \left( \frac{S^k}{k} \right) \alpha_D^k (1 - \alpha_D)^{S-k}. 
$$

(2.12)

The labeling of the populations is defined by $\alpha_{US} < \alpha_D$.

One motivation for using populations of neurons each with a particular value of sparsity arises from work by Attwell and Laughlin [24] and by Lennie [25]. There it is shown that to optimize the energy consumption by neurons transmitting a given amount of information, a particular value of sparsity is preferred, around a few percent. Even if other constraints besides energy consumption are important, the analyses suggest that sparsity is a parameter that could be adjusted (e.g., by evolution) to optimize the performance of a neural system. Therefore it is sensible
to propose models in which particular populations of neurons have particular values of sparsity. The assumption that all neurons in a population have exactly the same sparsity is an idealization. Allowing modest variations of sparsity within a population will not greatly change our results.

Another motivation for trying a multiple-population model is that recent experiments have detected multiple populations of neurons located in the various song-related regions of the brains of zebra finches [29]. These sets of neurons have strikingly different properties with regard to how often they spike.

This model successfully fits two of the four regions, with a large majority $\sim 95\%$ of the cells belonging to the ultra-sparse (US) population.

### 2.3.4 Beta Model

For the final model tested, we assume neuronal sparsity in the MTL follows a beta distribution. The PDF with parameters $a > 0$ and $b > 0$ is given by Eq (2.7) where $B(a, b)$ is the beta function. Motivation for choosing the beta distribution comes from its common usage as a distribution of probabilities, giving it wide application in Bayesian statistics [89]. The mean of the beta distribution is given by:

$$\langle \alpha \rangle = \frac{1}{B(a, b)} \int_{0}^{1} \alpha^a (1 - \alpha)^{b-1} = \frac{B(a + 1, b)}{B(a, b)} = \frac{a}{a + b}$$

(2.13)

Thus, for sparse codes, we would expect a low $\langle \alpha \rangle$, which would imply $a \ll b$. This can also be seen by examining Eq. (2.7). If $a \sim 0$, then the factor $\alpha^{a-1}$ is close to a power law with exponent $\sim -1$ implying that a huge proportion of cells have nearly zero sparsity. However, $a$ cannot equal zero exactly, as that would make it impossible to normalize the distribution.

Substituting (2.7) into (2.2) and evaluating the integral yields the beta-binomial distribution:

$$\epsilon_k (a, b) = \binom{S}{k} \frac{B(a + S, b + S - k)}{B(a, b)}.$$ 

(2.14)

Beta distributions with near power law behavior arise as limiting distributions in numerous “rich-get-richer” schemes. For example, both preferential attachment processes on growing networks, where nodes with a high degree are more likely to receive edges from newly added nodes [85,86], and Polya urn schemes [87], where the proportion of balls of a particular color grows whenever a ball of that color is
sampled from the urn, both yield beta distributions as \( t \to \infty \).

This model successfully fits all four regions.

### 2.4 Method of Maximum Likelihood

The method of maximum likelihood is a regression procedure for fitting a prediction from probabilistic model with parameters \( \tilde{\theta} \) to an observed data set. In general, after observing a particular data set during an experiment, we wish to find the model that produces the best prediction, i.e. we wish to maximize the probability that the model is parametrized by \( \tilde{\theta} \) given the observed data. Using Bayes’ theorem, we write:

\[
\operatorname{Prob}(\tilde{\theta} \mid \text{data}) = \frac{\operatorname{Prob}(\text{data} \mid \tilde{\theta}) \ast \operatorname{Prob}(\tilde{\theta})}{\operatorname{Prob}(\text{data})} \tag{2.15}
\]

The factor \( \operatorname{Prob}(\tilde{\theta}) \) is the prior probability that the model parameters take value \( \tilde{\theta} \), representing any previous knowledge. In cases where no prior knowledge of the model parameters is assumed, an unbiased prior is chosen, i.e. it is uniformly distributed across \( \tilde{\theta} \). In this case, maximizing Eq. (2.15) reduces to maximizing the factor \( \operatorname{Prob}(\text{data} \mid \tilde{\theta}) \) w.r.t. the parameters \( \tilde{\theta} \). This factor is called the likelihood function, \( L(\tilde{\theta}) \).

After deriving the likelihood function for the model, it is necessary to show that the \( n_k \) constitute “sufficient statistics” i.e. they fully represent the raw data. Finally, an approximation to the likelihood function is derived that applies in the case of sparsely responding units.

#### 2.4.1 The likelihood function

For this model, the likelihood function \( L(\tilde{\theta}) \) is the probability of observing the experimental outcome \( n_0, n_1, \ldots, n_S \) given model parameters \( \tilde{\theta} \).

\[
L(\tilde{\theta}) \equiv \operatorname{Prob}\left( \{ N_k = n_k \} \mid \tilde{\theta} \right) \tag{2.16}
\]

The likelihood function can be derived in a straightforward manner from single-unit responses. If the units are labeled \( j \in \{1, 2, \ldots, N\} \) and unit \( j \) responds to \( K_j \)
stimuli with probability $\epsilon_{kj}$, then the joint probability of the $K_j$ variables is

$$\text{Prob} \left( K_1 = k_1, \ldots, K_N = k_N \mid \tilde{\theta} \right) = \prod_{j=1}^{N} \epsilon_{kj} = \prod_{k=0}^{S} \epsilon_{k}^{n_k} \tag{2.17}$$

This probability represents one possible experimental outcome consistent with the $n_k$ values. There are many other possible outcomes that would yield the same result. Let $W\{n_k\}$ be the number of possible outcomes that would yield the same result. Let $W\{n_k\}$ be the number of possible outcomes that would yield the same result. Let $W\{n_k\}$ be the number of possible arrays of the $K_j$ variables that leave the $n_k$s fixed. Then our likelihood function is given by Eq. (2.17) multiplied by $W\{n_k\}$. Since $W\{n_k\}$ is independent of the model parameters, it does not need to be considered during maximization. However, for completeness, a derivation is given in the following paragraph.

$W\{n_k\}$ can be found by counting the combinations of bins one at a time starting with $n_0$. Given $N$ units, there are $\binom{N}{n_0}$ number of ways of choosing $n_0$ units that respond to zero stimuli. Of the remaining $N - n_0$ units, there are $\binom{N-n_0}{n_1}$ ways of choosing $n_1$ units that respond to only one stimulus. Continuing in this fashion to $n_S$, and multiplying the coefficients, we get the multinomial coefficient:

$$W\{n_k\} = \binom{N}{n_0} \binom{N-n_0}{n_1} \cdots \binom{n_{S-1} + n_{S}}{n_S} = \frac{N!}{\prod_{k=0}^{S} n_k!} \tag{2.18}$$

Then, the likelihood (Eq. (2.16)) function is

$$\mathcal{L} \left( \tilde{\theta} \right) = \frac{N!}{\prod_{k=0}^{S} n_k!} \prod_{k=0}^{S} \epsilon_{k}^{n_k} \tag{2.19}$$

Maximizing Eq. (2.19) w.r.t. $\tilde{\theta}$ gives the best fit $n_k^*$ that matches the data $n_k$. However, even after the best fit parameters are calculated, there is no guarantee that the fit matches the data. To test this, $\chi^2$ analysis is required.

### 2.4.2 Poisson Approximation

The data show that the probability of a neural response is much less than one. Thus we can usefully make approximations appropriate for the situation that $\epsilon_k \ll 1$ and $n_k \ll N$ for $k \geq 1$. Hence $\epsilon_0$ is close to unity, while $n_0$ is less than $N$ only by a
small fraction. Then

\[ \mathcal{L} \approx \prod_{k=1}^{S} e^{-N_{\epsilon_k}} \frac{(N_{\epsilon_k})^{n_k}}{n_k!}, \]  

(2.20)
i.e., a product of Poisson distributions for each of the non-zero \( k \) values. This is a standard result, but for completeness a derivation is given in Appendix A.

### 2.4.3 Sufficient Statistics

When performing likelihood analysis, special care must be taken in how the data is represented. If the raw data is given by the vector \( \bar{X} \), then the likelihood function for the raw data is given by

\[ \mathcal{L}_{\bar{X}} = \text{Prob} \left( \bar{X} \mid \theta \right). \]  

(2.21)

If we define some auxiliary statistics \( \tilde{N} \left( \bar{X} \right) \) that are functions of the raw data \( \bar{X} \), then we can calculate a different likelihood function for these new statistics,

\[ \mathcal{L}_{\tilde{N}} = \text{Prob} \left( \tilde{N} \left( \bar{X} \right) \mid \tilde{\theta} \right). \]  

(2.22)

Maximizing Eqs (2.21) and (2.22) will not necessarily yield the same parameters and thus infer different models. To avoid this problem, it is important therefore to check that if auxiliary statistics \( \tilde{N} \left( \bar{X} \right) \) are being used, that they will yield the same maximum likelihood parameters as with the raw data. Such statistics are called sufficient statistics.

In the case of this model, the raw data are the binary responses of each unit \( j \) to each stimulus \( s \), \( X_{js} \). The auxiliary statistics are the \( n_k \). The likelihood function of the \( n_k \) values is given by Eq(2.19). To show that the \( n_k \) are sufficient, we proceed to derive the likelihood function for the \( X_{js} \).

The \( X_{js} \) are a collection of \( NS \) Bernoulli random variables,

\[ X_{js} = \begin{cases} 
1, & \text{if neuron } j \text{ responds to stimulus } s, \\
0, & \text{if neuron } j \text{ does not respond to } s. 
\end{cases} \]  

(2.23)
The probability that \( X_{js} = 1 \) is given by \( \alpha_j \). Thus, the probability \( X_{js} = x_{js} \) is:

\[
\text{Prob} \left( X_{js} = x_{js} \mid \alpha_j \right) =
\begin{cases} 
\alpha_j, & \text{if } x_{js} = 1, \\
1 - \alpha_j, & \text{if } x_{js} = 0 
\end{cases} = \alpha_j^{x_{js}} (1 - \alpha_j)^{1-x_{js}}. \tag{2.24}
\]

Then, the joint conditional probability of the whole array \( X_{js} \) is given by

\[
\text{Prob} \left( \{ X_{js} = x_{js} \} \mid \{ \alpha_j \} \right) = \prod_{j,s} \alpha_j^{x_{js}} (1 - \alpha_j)^{1-x_{js}} = \prod_j \alpha_j^{k_j} (1 - \alpha_j)^{S-k_j} \tag{2.25}
\]

where \( k_j = \sum_s x_{js} \) is the number of stimuli to which unit \( j \) responds. The likelihood function of the raw data \( L_{\{X_{js}\}} (\hat{\theta}) \) is the total joint probability of the array of \( X_{js} \) variables, found by multiplying Eq. (2.25) by \( D(\alpha_1) \times \cdots \times D(\alpha_N) \) and integrating over \( \alpha_1 \ldots \alpha_N \):

\[
L_{\{X_{js}\}} (\hat{\theta}) = \int_0^1 d\alpha_1 D(\alpha_1) \ldots \int_0^1 d\alpha_N D(\alpha_N) \prod_j \alpha_j^{k_j} (1 - \alpha_j)^{S-k_j}
\]

\[
= \prod_j \int_0^1 d\alpha_j D(\alpha_j) \alpha_j^{k_j} (1 - \alpha_j)^{S-k_j} = \prod_j \frac{1}{\epsilon_{k_j}} \tag{2.26}
\]

In the last step, Eq. (2.2) is used to write the likelihood in terms of \( \epsilon_{k_j} \). Finally, we write \( L_{\{X_{js}\}} (\hat{\theta}) \) in terms of \( L_{\{n_k\}} (\hat{\theta}) \) using Eq. (2.19):

\[
L_{\{X_{js}\}} (\hat{\theta}) = \frac{1}{\prod_j \binom{S}{k_j} \sum_k \epsilon_k^{n_k}} \prod_k \frac{\epsilon_k^{n_k}}{W\{n_k\} \prod_j \binom{S}{k_j}} L_{\{n_k\}} (\hat{\theta}). \tag{2.27}
\]
Notice the prefactor is a combinatorial function of the data, and is independent of the model parameters $\theta$. Therefore, both $\mathcal{L}_{\{n_k\}}(\bar{\theta})$ and $\mathcal{L}_{\{X_{js}\}}(\bar{\theta})$ will be maximized at the same parameter values. Thus, the $n_k$ constitute sufficient statistics.

In another analysis of the same data [78], the statistics used were not sufficient.

### 2.5 Chi-squared Analysis

We will assess goodness of fit by using the following $\chi^2$ function:

$$\chi^2(k_{\text{max}}; n_1, n_2, \ldots) = \sum_{k=1}^{k_{\text{max}}} \frac{(n_k - N\epsilon_k)^2}{N\epsilon_k}.$$  \hspace{1cm} (2.28)

This corresponds to $-2 \ln \mathcal{L}$, to within an additive constant, when the Poisson distributions are replaced by Gaussian approximations near their peak. But this approximation is only suitable when $N\epsilon_k$ is reasonably much larger than one. So, in Eq. (2.28), we truncated the sum over $k$ to $k \leq k_{\text{max}}$. The restriction should be those values of $k$ such that the expected value of $n_k$ is bigger than one or two, i.e., where there are noticeable neural responses. Beyond $k_{\text{max}}$, there are very few neural responses, and therefore little information for fitting the parameters of the model.

After we have performed a maximum-likelihood estimate of the parameters of the model \(^1\), statistical theory [88] predicts a mean and standard deviation for $\chi^2$:

$$\chi^2 = N_{\text{dof}} = \sqrt{2N_{\text{dof}}}. \hspace{1cm} (2.29)$$

where the number of degrees of freedom, $N_{\text{dof}}$, is the number of data values used (i.e., the number of values of $k$ in the truncated sum) minus the number of parameters fitted (which will be 1, 2 or 3, in the particular implementations of the model that we use). If the value of $\chi^2$ falls much outside this range, that indicates that the model does not agree with the data.

\(^1\)Minimization of a suitably defined [90,91] $\chi^2$ is equivalent to maximizing likelihood.
2.6 Multiple Neuron Units

In the data that we analyze from [50], units were identified from electrode recordings using spike sorting techniques which cannot always distinguish individual neurons. Thus, the firing patterns reported for some units represent the aggregate firing of multiple neurons rather than for a single neuron. In fact, for the data we analyze, the multi-neuron units are roughly twice as abundant as the single-neuron units [1], and so it is crucial to consider the effect they have on the analysis.

The simplest version of our model ignores this distinction, and assumes that each recorded unit consists of only a single neuron. But, as we will explain the general principles of our methods apply perfectly well at the unit level. We will show how to transform from a neural-level model to a unit-level model, and we will see how to interpret numerical results of fits at the unit level in terms of single neuron properties. This will result in no change in our qualitative results, but a strengthening of our conclusions about the presence of a large number of very sparsely responding neurons.

First, the section begins by calculating the unit sparsity, \( \alpha_{\text{unit}} \), in terms of the constituent neuronal sparsities, \( \alpha_1, \ldots, \alpha_R \). Then, we proceed to derive the total probability that a unit consisting of \( R \) neurons responds to \( k \) stimuli, \( \epsilon_{\text{unit}}^{k,R} \), assuming the neuronal sparsities are each sampled from the same \( D_{\tilde{\theta}}(\alpha_r) \). Then, we calculate the probability that a unit selected at random (consisting of an unknown number of cells) responds to \( k \) stimuli, \( \epsilon_k^{\text{unit}} \), from,

\[
\epsilon_k^{\text{unit}} = \sum_{R=1}^{\infty} g(R) \epsilon_{k,R}^{\text{unit}}.
\]  
(2.30)

where \( g(R) \) is the fraction of units consisting of \( R \) neurons.

Finally, \( \epsilon_{k,R}^{\text{unit}} \) and \( \epsilon_k^{\text{unit}} \) are derived for each of the models considered in section 2.3. The remaining procedure for calculating the expected outcome, \( n_k^* \) (Eq. (2.3)) and maximizing the likelihood, \( \mathcal{L}(\hat{\theta}) \) (Eq. (2.19)) remains unchanged from the single-unit analysis above, only \( \epsilon_k^{\text{unit}} \) is used instead of the single unit probabilities, \( \epsilon_k \) given by Eq. (2.2). Also, it will be shown that the number of fitting parameters used in the multi-unit model does not change from the single-unit models, as \( g(R) \) is known.
2.6.1 Effective Sparsity

We now show that given our general multi-population model at the neuron level, a version of the model also applies at the unit level. That is, each unit has a sparsity, which can be calculated as a function of the sparsities of its constituent neurons, and there are populations of units with different values of sparsity.

Suppose first that a particular unit consists of $R$ neurons of known sparsities. Let $r = 1, \ldots, R$ label the neurons, and let $\alpha_r$ be the sparsity of neuron $r$. The unit’s sparsity, i.e., the probability that the unit responds to a presented stimulus, is given by:

$$\alpha_{\text{unit}} \equiv \text{Prob}(\text{response} \mid R, \alpha_1, \ldots, \alpha_R) = 1 - \prod_{r=1}^{R} (1 - \alpha_r).$$

(2.31)

Note that $\alpha_{\text{unit}} \geq \alpha_r$ for all $r$, consistent with the notion that the aggregated activity of multiple neurons should appear more responsive to stimuli compared to the activity of individual neurons.

Similar to the single-neuron unit, the conditional probability that the double-neuron unit responds to $k$ stimuli out of $S$ is given by the binomial distribution with probability parameter $\alpha_{\text{unit}}$.

$$P(K = k \mid \alpha_{\text{unit}}) = \binom{S}{k} \alpha_{\text{unit}}^k (1 - \alpha_{\text{unit}})^{S-k}.$$  

(2.32)

2.6.2 Single-Unit Response

The single-unit response, $\epsilon_{k,R}^{\text{unit}}$ is given above in Eq. (2.30). To derive the $\epsilon_{k,R}^{\text{unit}}$, the binomial distribution in Eq. (2.32) must be integrated over $D(\alpha_r)$ for each of the $r = 1 \ldots R$ neurons in the unit.

$$\epsilon_{k,R}^{\text{unit}} = \int_0^1 \ldots d\alpha_1 \int_0^1 d\alpha_R \left[ \prod_{r=1}^{R} D(\alpha_r) \right] \text{Prob}(K = k \mid \alpha_{\text{unit}})$$

(2.33)

In the case that $R = 1$ when the unit consists of a single neuron with sparsity $\alpha$, $\alpha_{\text{unit}} = \alpha$ and Eq. (2.33) simplifies to $\epsilon_k$ given by Eq. (2.2).
2.6.3 Multi-Unit Predictions

For our purposes, we restrict the analysis to the case where units consist of either a single neuron \( (R = 1) \) or two neurons \( (R = 2) \). This is an approximation, as in actual recordings, multi-units may consist of many more neurons. When \( R = 2 \), the effective sparsity (2.31) is given by

\[
\alpha_{\text{unit}} = 1 - (1 - \alpha_1)(1 - \alpha_2)
\]

(2.34)

where \( \alpha_1 \) and \( \alpha_2 \) are the sparsities of the two neurons. Eq. (2.30) simplifies to

\[
\epsilon_k^{\text{unit}} = p \epsilon_k + (1 - p) \epsilon_{k,2}^{\text{unit}}
\]

(2.35)

where \( p \) is the fraction of single-neuron units out of those recorded. For the data we analyze, \( p = 0.33 \) [1]. \( \epsilon_k \) is the single-unit response probability given by Eq. (2.2), and \( \epsilon_{k,2}^{\text{unit}} \) is the probability that a unit consisting of two neurons responds to \( k \) stimuli.

Finally, for \( R = 2 \), Eq. (2.33) simplifies to:

\[
\epsilon_{k,2}^{\text{unit}} = \int_0^1 d\alpha_1 \int_0^1 d\alpha_2 \left[ D_{\bar{\theta}}(\alpha_1) D_{\bar{\theta}}(\alpha_2) \times \left( \frac{S}{k} \right) \alpha_{\text{unit}}^k (1 - \alpha_{\text{unit}})^{S-k} \right]
\]

(2.36)

Given a sparsity distribution, \( D_{\bar{\theta}}(\alpha) \) and \( \alpha_{\text{unit}} \) defined in Eq. (2.34), we can calculate \( \epsilon_{k,2}^{\text{unit}} \) Eq. (2.36). Using \( \epsilon_{k,2}^{\text{unit}} \), along with the single-unit response probability \( \epsilon_k \), from Eq. (2.2); and the fraction of single-neuron units, \( p \); the total probability of a randomly chosen unit \( \epsilon_k^{\text{unit}} \) Eq. (2.35) can be computed. Finally, \( \epsilon_k^{\text{unit}} \) Eq. (2.35) is used to compute the expected outcome, \( n_k^* \):

\[
n_k^* = N \epsilon_k^{\text{unit}} \pm \sqrt{N \epsilon_k (1 - \epsilon_k^{\text{unit}})}.
\]

(2.37)

2.6.3.1 One Active Population

For the improved model that treats the recorded units as a mixture of single-neuron units and double-neuron units, the probability that a unit selected at random responds to \( k \) stimuli is (Eq. (2.35)), with \( \epsilon_k \) given by Eq. (2.10) and \( \epsilon_{k,2}^{\text{unit}} \).
calculated from Eqs. (2.36, 2.34, and 2.9) yielding:

\[
\epsilon_{k,2}^{\text{unit}} = \binom{S}{k} \alpha_k^{\text{unit}} (1 - \alpha_k^{\text{unit}})^{S-k}
\]  

(2.38)

with \( \alpha_k^{\text{unit}} = 1 - (1 - \alpha_D)^2 \). Thus, we can write the probability that a unit selected at random responds to \( k \) stimuli as:

\[
\epsilon_k^{\text{unit}} = p \binom{S}{k} \alpha_D^k (1 - \alpha_D)^{S-k} + (1 - p) \binom{S}{k} \alpha_k^{\text{unit}} (1 - \alpha_k^{\text{unit}})^{S-k}
\]  

(2.39)

The effect of adding a population of double-neuron units is to split the responses between single-neuron units with sparsity \( \alpha_D \) and double-neuron units with effective sparsity \( 1 - (1 - \alpha_D)^2 \). The double-neuron sparsity in this case exceeds \( \alpha_D \), indicating that units with multiple neurons will be observed to respond more frequently than those with single neurons, as expected. The predicted single-unit and double-unit activities can be seen in Figure 2.1, where the red curves indicate the single-unit contribution, and the yellow curves indicate the double-unit contribution. The overall expected response probability is the blue curve.

By comparing the overall multiunit probabilities given by the blue curve, with the purely single unit response probabilities given by the dotted line, we see that incorporating multi-neuron units into the model serves to spread out the response probabilities. In part (B) of Fig. 2.1, which plots the curves using expected parameter values based upon the experimental data in [1], the model incorporating multi-neuron units extends the tail. This result will be generally true of all models tested, and not just for the single population model.

### 2.6.3.2 Intermediate Model

When the naïve single population model is improved to include a silent population, the double-units come in three possible varieties. First, the double unit could consist of two silent neurons, in which case the unit also remains silent with \( \alpha_{\text{unit}} = 0 \). Secondly, it possible that the unit consists of one active neuron and one silent neuron, in which case \( \alpha_{\text{unit}} = \alpha_0 \). Finally, the double unit may consist of two active units, in which case, \( \alpha_{\text{unit}} = 1 - (1 - \alpha_D)^2 \).

The three possibilities can be derived using the \( D_{\hat{\theta}}(\alpha) \) for the intermediate
model (Eq. (2.5)) and substituting into (2.44). Solving the integral, we get the double-unit response probability for the model with one active and one silent population:

\[
\epsilon_{k,2}^{\text{unit}} = (1 - f_D)^2 \delta_{k,0} + 2f_D (1 - f_D) \binom{S}{k} \alpha_D^k (1 - \alpha_D)^{S-k} + f_D^2 \binom{S}{k} \alpha'^k (1 - \alpha')^{S-k}
\]

where \( \alpha' = 1 - (1 - \alpha_D)^2 \) and \( \delta_{k,0} \) is the Kronecker delta. Eq. (2.40) represents solely the activity of the double-units. When the single units are included in the population, the overall probability that a randomly selected unit responds to \( k \) stimuli is given by:

\[
\epsilon_k^{\text{unit}} = p \left\{ (1 - f_D) \delta_{k,0} + f_D \binom{S}{k} \alpha_D^k (1 - \alpha_D)^{S-k} \right\} + (1 - p) \left\{ (1 - f_D)^2 \delta_{k,0} + 2f_D (1 - f_D) \binom{S}{k} \alpha_D^k (1 - \alpha_D)^{S-k} + f_D^2 \binom{S}{k} \alpha'^k (1 - \alpha')^{S-k} \right\}
\]

Figure 2.1. Figures indicating the effect of including multi-neurons units on the one-population model. The blue circles indicate the probability that a randomly selected unit responds to \( k \) stimuli, given by Eq. (2.39). The red squares and the yellow diamonds represent the weighted binomial curves calculated from the first and second terms of Eq. (2.39) respectively, indicating the responses of single-neuron units and double-neuron units. The black dashed line indicates the probabilities calculated by assuming only single units (2.10). The plot (A) on the left is for \( S = 100 \) stimuli shown, which is what is shown in a typical experiment of the type analyzed. The plot (B) is with an exaggerated \( S = 2000 \) images presented. For both plots, \( p = 0.33 \) as reported in [1] and \( \alpha_D = 0.005 \).
2.6.3.3 Two Active Populations

The multi-unit results for two active populations with sparsities $\alpha_{US}$ and $\alpha_D$ take on a similar form as with the intermediate model above. The double units in this model also take three possible forms: 1. both neurons have sparsity $\alpha_{US}$; 2. both neurons have sparsity $\alpha_D$; and 3. one neuron is from the US population while the other is from the D population.

As with the intermediate model, we calculate the double-unit response probability by substituting the sparsity distribution (Eq. (2.6)) and substituting into (2.36).

\[
\epsilon_{k,2}^\text{unit} = (1 - f_D)^2 \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{US}^k (1 - \alpha_{US}^\prime)^{S-k} + 2f_D (1 - f_D) \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{D}^k (1 - \alpha_D^\prime)^{S-k} + f_D^2 \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{D}^k (1 - \alpha_D^\prime)^{S-k} \tag{2.42}
\]

where $\alpha_D^\prime = 1 - (1 - \alpha_D)^2$, $\alpha_{US}^\prime = 1 - (1 - \alpha_{US})^2$, and $\alpha'' = 1 - (1 - \alpha_D) (1 - \alpha_{US})$. Including the single units, give the overall response probability;

\[
\epsilon_k^\text{unit} = p \left\{ (1 - f_D) \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{US}^k (1 - \alpha_{US}^\prime)^{S-k} + f_D \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{D}^k (1 - \alpha_D^\prime)^{S-k} \right\} \\
+ (1 - p) \left\{ (1 - f_D)^2 \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{US}^k (1 - \alpha_{US}^\prime)^{S-k} + 2f_D (1 - f_D) \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{D}^k (1 - \alpha_D^\prime)^{S-k} + f_D^2 \left( \begin{array}{c} S \\ k \end{array} \right) \alpha_{D}^k (1 - \alpha_D^\prime)^{S-k} \right\} \tag{2.43}
\]

Plots of $\epsilon_k^\text{unit}$ in the two-population model are shown by the blue curves in Fig. 2.2. The red curves indicate the single unit contributions to the response (the two terms weighted by $p$ in Eq. (2.41)) while the yellow curves indicate the double-units. Similar to the case of the one-population model, the incorporation of multi-units into the two-population model increases the weight of the tail under expected experimental conditions (Fig. 2.2 (B)) reported in the data we seek to fit [1].

It will be shown that the goodness of fits are largely preserved by incorporating
Figure 2.2. Figures indicating the effect of including multi-neurons units on the two-population model. The blue circles indicate the probability that a randomly selected unit responds to \( k \) stimuli, given by Eq. (2.41). The red squares and the yellow diamonds represent first and second terms of Eq. (2.41) respectively, indicating the responses of single-neuron units and double-neuron units. The black dashed line indicates the probabilities calculated by assuming only single units 2.12. The plot (A) is for \( p = 0.33 \), with \( S = 100 \) images presented, \( \alpha_D = 0.02 \), \( \alpha_{US} = 0.001 \), and \( f_D = 0.2 \). The plot (B) on the right is the same as (A), only with an exaggerated \( S = 1000 \). Here the different behaviors of the single and double units is made apparent. The single unit response clearly indicates two peaks, one for the D population centered at \( k \sim 20 \) and one for the US population near \( k = 0 \). The double unit response shows three peaks: US-US at \( k \sim 0 \), US-D at \( k \sim 20 \), and D-D which is barely visible at \( k \sim 40 \).

the multi-units, however the parameter values of the best fits indicate a sparser code. This is to be expected, as multi-units respond more frequently than single units. The nature of the two-population, multi-unit fits is discussed in detail in Sec. 3.5.

2.6.3.4 Beta Distribution

In the beta model, each neuron in a double unit has sparsity sampled from a beta distribution (Eq. (2.7)). The total probability for the double unit response \( \epsilon_{k,2}^{\text{unit}} \) is found by integrating Eq. (2.44) for each constituent neuron:

\[
\epsilon_{2,k}^{\text{unit}} = \int_0^1 d\alpha_1 \int_0^1 d\alpha_2 \left[ D_{a,b}(\alpha_1) D_{a,b}(\alpha_2) \left( \frac{S}{k} \right)^{\alpha_k^{\text{unit}}(1 - \alpha_{\text{unit}})^{S-k}} \right] \tag{2.44}
\]

To evaluate the integral, we first note that:

\[
(1 - \alpha_{\text{unit}})^{S-k} = [(1 - \alpha_1)(1 - \alpha_2)]^{S-k} \tag{2.45}
\]
Then, we expand the factor $\alpha unit_k$ in (2.44) using the binomial theorem:

$$\alpha unit_k = (\alpha_1 + \alpha_2 (1 - \alpha_1))^k = \sum_{j=0}^{k} \binom{k}{j} \alpha_1^j [\alpha_2 (1 - \alpha_1)]^{k-j}$$  (2.46)

Using these results and Eq. (2.7), we can write Eq. (2.44) as a sum of separable integrals

$$\epsilon unit_{2,k} = \frac{1}{[B(a, b)]^2} \binom{S}{k} \sum_{j=0}^{k} \left\{ \binom{k}{j} \right\} \times \int_0^1 d\alpha_1 \alpha_1^{a-j} (1 - \alpha_1)^{b-1+S-j} \times \int_0^1 d\alpha_2 \alpha_2^{a-k+j} (1 - \alpha_2)^{b-1+S-k}$$  (2.47)

The integrals can now be evaluated as beta functions

$$\epsilon unit_{2,k} = \frac{1}{[B(a, b)]^2} \binom{S}{k} \sum_{j=0}^{k} \left\{ \binom{k}{j} \right\} B(a+j, b+S-j) B(a+k-j, b+S-k) \quad \text{(2.48)}$$

Finally, the total probability that a randomly chosen unit responds to $k$ stimuli is given by:

$$\epsilon unit_k = \frac{p}{B(a, b)} \alpha^{a-1}(1 - \alpha)^{b-1} + \frac{1-p}{[B(a, b)]^2} \binom{S}{k} \sum_{j=0}^{k} \left\{ \binom{k}{j} \right\} B(a+j, b+S-j) B(a+k-j, b+S-k) \quad \text{(2.49)}$$

The plots in Fig 2.3 show Eq. (2.49) as the blue curve. The single unit and double unit activity is shown by the red and yellow curves, respectively. The dotted line is the probability predicted assuming only single units. As with the one and two population models, incorporating the multi-unit activity into the beta model spreads the response distribution.

However, there is a key distinction between the population models and the beta models: the beta model preserves its shape as the number of stimuli increases,
Figure 2.3. Figures indicating the effect of including multi-neurons units on the beta model. The blue circles indicate the probability that a randomly selected unit responds to \( k \) stimuli, given by Eq. (2.49). The red squares and the yellow diamonds represent first and second terms of Eq. (2.49) respectively, indicating the responses of single-neuron units and double-neuron units. The black dashed line indicates the probabilities calculated by assuming only single units 2.14. The plot (A) is for \( p = 0.33 \), with an exaggerated \( S = 1000 \) images presented, \( a = 0.1 \), and \( b = 60 \). The plot (B) on the right is the same as (A), only with \( S = 100 \), which is the number used in the data we fit.

whereas the shape of the probability responses are not preserved in the population models. We can see this by comparing (A) and (B) in Fig. 2.3. While (B) is plotted assuming \( S = 100 \) stimuli are shown, (A) shows the plots for when a much larger number \( S = 1000 \) are shown. However, both plots possess the same shape.

Now, if we look at the population plots shown in Figs. 2.1 and 2.2, the shape of the prediction changes dramatically as the stimuli are increased. At experimentally tested \( S = 100 \), the beta and two-population predictions are difficult to distinguish. This is why, in the following chapter, both the two-population model and the beta model will successfully fit the same data, despite being radically different distributions.

In a hypothetical experiment showing thousands of stimuli to the patients, the predictions should be considerably different, allowing us to determine the correct model.
Chapter 3  
Model Fits

3.1 Introduction

The goal of this chapter is to apply the models developed in the Chapter 2 to the unit recordings from the human MTL presented in Mormann et al. [50].

First, the one-population and intermediate models will be shown to produce bad fits to this data, ruling out the possibility that the active neurons of the MTL possess a uniform sparsity. Then, the two-population and beta models will be shown to successfully fit the hippocampus and entorhinal cortex responses, according to the $\chi^2$ values, while the beta model alone successfully fits the amygdala and the parahippocampal cortex.

The improved model that includes multi-neuron units will also be tested against the data, resulting in similar $\chi^2$ values as with the simpler models, but predicting significantly sparser distributions. The successful fits indicate an extremely sparse binary code in all regions reported. For all fits, likelihood maximization was performed numerically in Mathematica, and we used the exact formula (2.19) for the likelihood, without any approximation such as (2.20).

The data [50] being analyzed come from recordings of single neuron activity in four regions of the human MTL, the hippocampus (Hipp), the entorhinal cortex (EC), the amygdala (Amy), and the parahippocampal cortex (PHC). Altogether, 1194 neurons/units were detected in the hippocampus, 844 in the entorhinal cortex, 947 in the amygdala, and 293 in the parahippocampal cortex, accumulated over patients. During the experiment, the patients were shown a randomized sequence of images of famous individuals, landmarks, animals, and objects. For each session,
Table 3.1. Number of neurons $n_k$ responding to $k$ images as reported by [50] in four MTL regions.

<table>
<thead>
<tr>
<th></th>
<th>$n_0$</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
<th>$n_4$</th>
<th>$n_5$</th>
<th>$n_6$</th>
<th>$n_7$</th>
<th>$n_8$</th>
<th>$n_9$</th>
<th>$n_{10}$</th>
<th>$n_{11}$</th>
<th>$n_{12}$</th>
<th>$n_{13}$</th>
<th>$n_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hipp</td>
<td>1019</td>
<td>113</td>
<td>30</td>
<td>17</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EC</td>
<td>761</td>
<td>45</td>
<td>15</td>
<td>9</td>
<td>4</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Amy</td>
<td>842</td>
<td>61</td>
<td>17</td>
<td>15</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PHC</td>
<td>244</td>
<td>13</td>
<td>11</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The patients were shown on average 97 images. Histograms of the number of units giving above-threshold responses for particular numbers of stimuli were then produced. The values are given in Table 3.1, which we read from graphs in Fig. 3 of [50].

The results of the fits are presented in Table 3.2, we tabulate the numerical results of the fits of each model in each region. In Figs. 3.1–3.4, we show the results graphically; each figure shows the results of the three fits for a single region.

### 3.2 Failure of One-Population and Intermediate Models

#### 3.2.1 One-Population fits

We assess the goodness of the fit using $\chi^2$ defined in Eq. (2.28). The enormous values of $\chi^2$ — see Table 3.2(a) — show that the fits are extremely bad, given the expected range, Eq. (2.29).

The nature of the bad fit is seen in the plots in the top row of each of Figs. 3.1–3.4. Here we compare the data with the expectation values of $n_k$ given by the best fit. Because the value of $n_0$, the number of non-responding neurons, is much larger than for the other $n_k$, we show two versions of the plot. The right-hand plots have the $k = 0$ bin omitted and have a changed vertical scale, to better exhibit the other bins.

The data exceed the best fit in the $k = 0$ bin in all 4 regions, they undershoot it in the $k = 1$ bin, and are substantially non-zero in several bins with $k \geq 3$, where the fit is close to zero. We have a choice. If the sparsity is large enough to give a sufficient number of cells that respond to multiple stimuli, then, compared with data, too few cells are non-responsive. If instead, the sparsity is low enough to
Figure 3.1. Comparison of data from the hippocampus with fits for the number of neurons \( n_k \) that respond to \( k \) stimuli. The red circles are the experimental values and blue dots connected by lines are the model predictions for the expectation values of \( n_k \). The blue error bars are the models’ predictions for the one-standard-deviation variation of experimental results on repetition of the experiment. The top plots (A) are for fits with the one-population model. The plots (B) are for the intermediate model with one active and one silent population. The fits (C) are for the model with two active populations, and finally the bottom plots (D) are for the continuous beta model. The right-hand plots differ from the left-hand plots only by having a larger vertical scale, to show more clearly the \( k \geq 1 \) bins.
Table 3.2. Results of the fits of the four models. For each model, we give the values and uncertainties of the model’s parameters for each brain region, and we give the $\chi^2$ measuring the goodness of fit.

reproduce the number of non-responsive cells (i.e., $n_0$), then there is much too small a probability for multiple responses. In either case, the model cannot reproduce all the data.

To see the one-population model compared with the beta model, see Fig. 4.2.
3.2.2 Intermediate Model Fits

The results of a maximum-likelihood fit are shown numerically in Table 3.2(b), and graphically in the middle row (B) of each of Figs. 3.1–3.4. The fits are improved, but the $\chi^2$ values are still substantially too large for a good fit. Notice how the silent population contains by far the majority of the neurons in all four regions.

The pattern of deviations between data and model is now an excess for the data in the $k = 1$ bins and a deficit at $k = 2$. That is, the number of cells responding exactly once is substantially higher compared with the extrapolation of the numbers of cells with multiple responses. This indicates that a better model would be to replace the silent neural population by a slightly active population. To fit the data, this population must have a very small sparsity, so that it predominantly gives contributions to the $k = 0$ and $k = 1$ bins only.

3.3 Two-Population Fits

The results of a maximum-likelihood fit are shown numerically in Table 3.2(c), and graphically in the bottom row (C) of each of Figs. 3.1–3.4. The fits are much improved. For both the hippocampus and the entorhinal cortex, we have good fits, with the model being consistent with the data. The fit in the amygdala is less reliable and the model poorly fits the data in the parahippocampal cortex. In all cases, the ultra-sparse population is in the vast majority, around 90% or more, while at the same time having a very small sparsity, $10^{-3}$ or smaller. Thus each neuron in the ultra-sparse population responds on average to at most about 0.1 stimuli in a session. We only see the effects of the ultra-sparse population because the data are from a large number of neurons. In contrast, the remaining few percent of neurons in the other population typically respond to several stimuli in each session.

For the parahippocampal cortex (PHC), a different or more general model is clearly needed. We observe that the functionality of the PHC is much different than that of the hippocampus and the entorhinal cortex, so it is not surprising that its neural coding properties should be different. The hippocampus is the classical locus of episodic memory storage, and the entorhinal cortex is its main source of input (and output).

An alternative view of the fit is shown in Fig. 3.5. Here we show how the neural
responses are predicted by the model to arise from the different populations. The bottom parts of the bars, shaded gray, show the expectation values for the part of $N_k$ coming from above-threshold responses by neurons in the D population. Stacked above these are open bars, showing the contribution from the US population. In the bins with more than one response, i.e., $k \geq 2$, almost all the responses are from the D population, with only a small contamination from the ultra-sparse population, primarily at $k = 2$. In contrast, in the $k = 1$ bin, there is a relatively small fraction of responses from the D population, from the tail of a distribution with its peak at several responses. The majority of the $k = 1$ bin is from the ultra-sparse population. However, this represents only the tip of the iceberg, so to speak. The vast majority of the ultra-sparse neurons give no above-threshold responses; they appear in the $k = 0$ bin, which is much too tall to be shown in Fig. 3.5.

The sparsity and fraction for the D population can be determined from the $k \geq 2$ bins, i.e., with the exclusion of the $k = 1$ bin. There are several bins involved, so the shape of the distribution of $N_k$ from a single sparsity fit is confirmed, as can be seen from the lowest plots in the right-hand column of Figs. 3.1–3.4, certainly for the HC and EC regions. Extrapolating the fit for the D population to $k = 1$ falls far short of the data, by a factor of at least 5. This then determines that there is an ultra-sparse population, whose average sparsity is determined to a good approximation by the excess in the $k = 1$ bin relative to the total number of non-D neurons:

$$\alpha_{US} \approx \frac{N_1(\text{excess above extrapolation})}{N(1 - f_D)S}. \quad (3.1)$$

Our actual best fit allows for the contamination of the bins of higher $k$ by the ultra-sparse population. The existence and size of the ultra-sparse population is determined by the large excess of the measured value of $N_1$ compared with the extrapolation from the bins of larger $k$, whose relative sizes correspond to a sparsity of a few percent.

The confidence intervals on the parameters displayed in Table 3.2 are calculated according to the procedure detailed in Appendix D.
### 3.4 Beta Fits

All four regions are successfully fit by both the beta model containing single units and the improved model containing a mixture of single and double neuron units ($\chi^2$ values given in Table 3.2 (d)). This suggests that the data is consistent with the notion that the sparsity of human MTL neurons follows a beta distribution, given in Eq. (2.7), with $a < 1$ and $b > 1$. The sparsity distribution is far from uniform, consistent with [73].

Including the double units in the model had little effect on the goodness of fit, though the value of the $a$ parameter is brought closer to zero (3.2). This results in a lower mean sparsity compared with assuming only single neuron units.

Plots of the best-fit sparsity distributions from the mixture model are given in figure [3.6]. The means of the distribution in each of the regions are: $\langle \alpha \rangle = 1.6 \times 10^{-3}$ in Hipp; $1.5 \times 10^{-3}$ in EC; $1.5 \times 10^{-3}$ in Amy; $4.0 \times 10^{-3}$ in PHC. In each region, $a$ is close to zero, producing a near-divergence in the neuronal sparsity distribution as $\alpha \to 0$. This indicates a large population of extremely sparse neurons, consistent with the results of [73]. For these fits, roughly 95% of the MTL neurons falls within the power-law regime, as is shown by the linear region of the log-log plots in figure 3.6.

The sparsity distribution in the PHC differed quantitatively from the fits in the other three regions. Firstly, the model predicts that neurons of the PHC has the highest mean sparsity, $\langle \alpha \rangle = 4 \times 10^{-3}$, compared to the other three regions which have mean sparsities clustered near $1.5 \times 10^{-3}$. Also, the tail of the sparsity distribution extends further for the PHC than it does for the other three regions, as shown in figure [3.6]. This is consistent with the observation that the selectivities of neurons tend to increase as information is processed at higher and higher stages. In the MTL the PHC has been shown to possess the lowest latency compared to the EC, Hipp, and Amy [50].

The success of the beta distribution leads to further questions about the MTL code. What is a possible generating mechanism? What is the utility of having power-law like behavior and a highly skewed distribution? These questions will be explored in detail in sections 4.3 and 4.2.4.

To see the beta model juxtaposed with the single population model, see Fig. 4.2.
Table 3.3. MLE values and $\chi^2$ values for a multi-unit model with two neural populations. Compared with the parameters fit using the single-unit model, in Table 3.2 (c), the multi-unit model shows a lower sparsity value for $\alpha_{US}$ indicating a sparser code as discussed in Sec 2.6.

### 3.5 Multi-Unit Results

The results presented thus far have been under the assumption that all recorded units are composed of a single neuron. However, it is known that some units are in fact composed of multiple neurons [1,50]. To gain an idea of the effect of such multi-units on our fits and of our conclusions about neural properties, we apply the general analysis from Sec. 2.6.

#### 3.5.1 Multi-unit fits: two population model

If only single-neuron and double-neuron units existed, i.e., if only $R = 1$ and $R = 2$ occur, then the total number of populations at the unit level would be 5, and these would appear in the formula (2.30) for $\epsilon_{\text{unit}}^k$. If larger multi-units occur, there are even more populations of units, each with its particular sparsity. This seems like a very complicated situation, but provides no issue of principle in the MLE of the parameters of the populations at the neural level, except that there is little data about the exact distribution of the number of neurons in a unit.

However, as we will see in more detail shortly, considerable simplifications occur because the vast majority of neurons are extremely sparsely firing. This property is reflected at the unit level, and we will see that from a two-population model at the neural level, the unit-level data are reasonably accurately given by a two-population model at the unit level. This justifies a posteriori the success of a two-population model applied to unit level data, and one can see how to relate properties of the neural populations to properties of the unit populations. The reasons come from the two most common kind of unit. The most common situation is that all the
neurons in a unit are all ultra-sparse, so that the unit itself responds ultra-sparately, typically at most one neuron at a time. The second most common situation is where exactly one of the neurons in a unit is in the D population. Then by far the most common response from the unit is due to the single D neuron.

We then used this in the formalism that we have already set up. The resulting fitted values of the parameters and the $\chi^2$ values are shown in Table 3.3 and, for the case of the hippocampus, the plots of the predicted and recorded $n_k$ are shown in Fig. 3.7. Compared to the values fitted for the original two-population model, Table 3.2(c), the $\chi^2$ did not change greatly, although the goodness of fit is somewhat improved, notably in the amygdala and PHC.

Thus the multi-unit model fits all four regions at least as well as the basic two-population model, without any extra fitted parameters. However, the values of some of the parameters did change. Most notably, the estimate of $\alpha_{US}$ decreased by about 40%. The earlier value is simply a weighted average of the effect of single units containing one US neuron with sparsity $\alpha_{US}$ and double units containing two US neurons, with unit sparsity $2\alpha_{US}$. The fitting of the US population is determined primarily from the $n_0$ and $n_1$ bins, so the response data by itself cannot significantly determine the existence of these two populations of units; it just gives a weighted average of the sparsities.

In contrast, the value of $\alpha_D$ is only slightly lower than in the earlier fit; this is because most of the relevant data concerns units that have one D neuron. Given the sparseness at the unit level, as in Table 3.2(c), which is tied to measured data, the neural sparsity must be less when there are multi-units.

As to the population fractions, the value $f_D$ is reduced by a factor of about two thirds. This is because for a given $f_D$ at the neural level, there are two chances of having a D neuron in a double unit. Thus the effective value of $f_D$ at the unit level is the following weighted average:

$$f_D(1 - p) + 2f_Dp = f_D(1 + p), \quad (3.2)$$

which determines the difference between the values of $f_D$ in Tables 3.2(c) and 3.3 fairly well, given the value $p = 0.34$ that we deduced from Ref. [50].
Table 3.4. Parameter values and $\chi^2$ values for the multi-unit fits in the beta model. Compared with the single-unit beta fits in Table 3.2 (d), the multi-unit fits exhibit improved $\chi^2$ values in most of the regions, and a smaller exponent, $a$.

3.5.2 Beta Multi-unit Fits

The fits of the beta model including double-units are shown in Fig. 3.8, with the parameters given in Table 3.4. The quality of fits slightly improved and decreased the $a$ parameter, which governs the near-divergence at zero sparsity. This serves to decrease the average sparsity of the four regions of the MTL (see section 2.3.4).

Including the double units in the model had little effect on the goodness of fit, though the value of the $a$ parameter is brought closer to zero (3.2). This results in a lower mean sparsity compared with assuming only single neuron units.

3.5.3 Sparsening Effect

We see that allowing for multi-units has actually strengthened our conclusion that there is a large fraction of extremely sparsely responding neurons. Effectively the existence of multi-units has diluted the effect in the data relative to the situation at the neural level. The original two-population and beta fits in Table 3.2(c) characterize the measured data at the unit level. At the neural level, as supported by Table 3.3 and Table 3.4, the fraction of neurons with nearly zero sparsity grows. This qualitative result is independent of the exact numbers of multi-units and the distribution of numbers of neurons in a unit.

3.5.4 Multi-units did not salvage one-population model

The inclusion of double units did not improve the predictions of the single population or the intermediate models. Both models still failed, even though the double units and the single units act similarly to two different populations. However, the effective sparsity of the double-units is only twice as large as the sparsity of the single-unit population, as can be seen in Sec 2.6.3.1. Comparing this with the fits from the full
two-population model, where the two populations have hugely different sparsity values ($\alpha_{US} \sim 10^{-3}$ vs $\alpha_{D} \sim 10^{-2}$), we see that the two populations arising from single-vs-double units are insufficiently different to account for the large tail in the data.

### 3.6 Comparison

We can see that the results of fitting the single-population model with its single sparsity were intermediate sparsities compromising between the extremes of the two populations in the full model. The single-population model therefore incorrectly represents the actual neural sparsity. Our fitted values of sparsity in Table 3.2(a) roughly match those found by Waydo et al. [78] in their fit of a pure one-population model to similar data.

In the intermediate model, with a set of exactly silent neurons, the fit for the responsive neurons is qualitatively similar to the D neurons in the full model: a sparsity of a percent to a few percent and a minority abundance. Relative to the full model, the value for the active population’s sparsity is still biased downwards, while its fractional abundance is biased upwards; these properties give a compromise between the effects of the two populations of neurons in the full model.

Merely introducing extra parameters increases the goodness of fit, but only by an expected decrease of one unit in $\chi^2$ per parameter, as in Eq. (2.29). So the improved fits from adding an extra population are highly significant. In all cases, the $p$-values for the poor fits for the first two models are well below 0.001, from standard plots or tables for the $\chi^2$ distribution.

Both the beta model and the two-population model produced successful fits in the hippocampus (Fig. 3.1 (C) and (D)) and the entorhinal cortex (Fig. 3.2, (C) and (D)), but the beta model had somewhat improved fits in the amygdala and parahippocampal cortex (Figs. 3.3, 3.4). Furthermore, the beta model utilizes two fitting parameters, $a$ and $b$ while the two-population model uses three, $\alpha_D$, $\alpha_{US}$, and $f_D$. The similarities of the predictions of these two radically different models originates from the relatively low number of stimuli used during the experiment ($S \sim 100$). In experiments with larger numbers of stimuli ($S \sim 1000$), the models would be clearly distinguished, as discussed in section 2.6.3.4.
Figure 3.2. The same as Fig. 3.1, but for the entorhinal cortex.
Figure 3.3. The same as Fig. 3.1, but for the amygdala.
Figure 3.4. The same as Fig. 3.1, but for the parahippocampal cortex.
Figure 3.5. Plots of neuron responses predicted by the three-parameter maximum-likelihood fits to data in four regions of the MTL. The plots are of number of neurons as a function of number of responses. The shaded bars represent neurons in the almost-silent population while the open bars correspond to the distributed population. The red circles indicate the experimental values.
Figure 3.6. Sparsity distributions across neurons in each region predicted by the multi-unit model consisting of single units and double units. The plots are shown using log-log axes to indicate power law behavior at low sparsity. The shaded region indicates the upper 5% tail of the sparsity distribution, i.e. the model predicts 95% of the neurons in each region have a sparsity left of the shaded area. The verticle dotted lines indicate the mean sparsities in each region.

Figure 3.7. Fits of the two-population multi-unit model applied to the hippocampus.
Figure 3.8. Fits of the beta multi-unit model applied to the hippocampus.
Chapter 4  |  Coding Implications

4.1 How Sparse is Sparse?

The primary result is that in the hippocampus\(^1\) (and other areas of the MTL), a vast majority (90% or higher) of neurons respond ultra-sparingly to stimuli in the class presented — around one in a thousand stimuli, or even less. Those neurons that respond more readily, i.e., to a few percent of the stimuli, comprise a rather small fraction (several percent) of the cells.

We have devised methods that treat the measured neurons and stimuli as statistical samples, and allow the deduction of properties of the ultra-sparse population even though these neurons respond on average to less than a tenth of a stimulus out of the approximately 100 used to obtain the data analyzed [50].

Our results confirm and substantially strengthen results found by Collins and Jin in [72]. There, earlier data from [1] was used that only provided values for the number of units with one response, \(N_1\), the number with two or more responses, \(N_{\geq 2}\), and for the average number of responses from responsive units. They ruled out a one-population model, and fit the three parameters of the two-population model from the three reported summary statistics. But there was therefore no test of goodness of fit for the 3-parameter model. Instead the shape of the \(N_k\) distribution was a prediction, which is successfully tested in this paper, for the case the hippocampus and entorhinal cortex, at least.

In the paper by Magyar and Collins [73], we also give a more systematic account of the statistical methods, and have a breakdown by brain region, using newer data.

\(^1\)In the paper reporting the data we use, it is not stated which hippocampal region the recorded cells belong to.
Although the exact parameters of the fits are a little different, the main picture is confirmed and tested. The differences could be accounted for by differences in the subjects and of detailed experimental procedures and by our different treatment of multi-units.

### 4.1.1 Small sparsity values

Although our model provides a good fit to the data (at least in the hippocampus and entorhinal cortex), it should not be supposed that the model gives an exact characterization of neural responses to stimuli.

The first issue is simply that if we replaced one particular neural population of a particular sparsity $\alpha$ and fractional abundance $f$ by several populations with sparsities not too far from the original single value, and with a total abundance summing to $f$, the result would not be very distinguishable from the original case. This is simply because the probability distribution (2.1) for $k$ in a single population is substantially broader than the delta function $\delta(k/S - \alpha)$ that would apply in the limit of an infinite number of stimuli. This issue is particularly notable for the ultra-sparse population, which primarily gives contributions only to the $k = 1$ bin. Thus our measurement of a sparsity for the ultra-sparse population is really of a weighted average of the sparsities of the ultra-sparlessly responding neurons.

What is properly deduced is that there are relatively few neurons that respond with a sparsity of a few percent, and a much larger number that respond much more sparsely.

### 4.1.2 How close are MTL neurons to GM cells?

It has been suggested that the neurons under discussion are concept cells [54]; each cell responds to the presence, in some sense, of a particular concept in the current stimulus. One part of the motivation for this is that the responses often appear to be genuinely conceptual; for example, a cell might respond (within the experimental data) only to stimuli involving a particular person, e.g., to multiple different pictures of the person, even in disguise, to the written name of the person, and to the spoken name [1,51].

It is tempting to say that because a cell consistently responds to a variety of very different stimuli related to a particular person, e.g., Jennifer Aniston, that the
cell actually represents Jennifer Aniston, i.e., that its firing above threshold codes that the concept of Jennifer Aniston is being processed or has been detected. This is the simplest version of the concept cell idea [54].

In this section, we make some suggestions about how our results quantitatively impact this issue. We label the subject that of the semantics of the cells, i.e., of what meaning should be attached to their responses.

First, the results do strengthen the basic case for conceptual cells, by quantifying how sparse the responses typically are. As explained in Sec. 1.2.1, our results are fits to the screening session data. The consistency between stimulus presentations and the results of the testing sessions support the conceptual nature of the responses.

We propose the following questions to clarify the concept-cell issues: One is whether the concept involved is in fact the obvious one, e.g., Jennifer Aniston for stimuli involving her. Next is whether the concept is one that in some sense is present in the current stimulus, be it visual or auditory. Our third question is whether the neural conceptual representation is strictly local, i.e., whether the above-threshold firing of one of these cells codes only a single concept. Finally, we ask whether the representation for only one concept is active at a given time or whether the representations of multiple concepts are (more-or-less) simultaneously active, on a time scale of a few hundred milliseconds. Discussions of concept representations, including [54], often appear to us to unnecessarily presuppose particular answers to these questions.

First, it appears necessary that a concept, like Jennifer Aniston, is coded in the subject’s brain. But the reported cell can be a downstream consequence. For the reported Jennifer Aniston cell, it was later found [54] that the cell also responded to a picture of Lisa Kudrow, a co-star of Aniston’s in a television series. Thus, the concept coded by the neuron might not be the obvious one, but the neuron might instead code the activation of a related episodic memory, for example.\(^2\)

Once one allows that downstream concepts are activated, one should expect that multiple concepts are simultaneously active, with perhaps only one being selected at a given time for conscious attention.\(^3\)

\(^2\)That individual episodic memories may be among the concepts involved (for some definition of “concept”) is suggested because of the well-known role of the hippocampus in the episodic memory system.

\(^3\)This is essentially the same property that computer search engines have, which we can regard as a kind of associative memory system that can give multiple responses to a search cue.
With regards to our second question, we know from everyday experience that recall of memories can be based not directly on a current stimulus but on cues generated by internal cognition. In general, this removes an obligatory link between stimulus properties and conceptual neural responses. For the experimental measurements under discussion, this issue need not be important, because the experimental protocol was explicitly designed to have the subjects’ attention focused on the stimuli.

As to whether or not the representation is local, normally a dichotomy is made simply between local representations and distributed representations. In the case of distributed representations, even when they are sparse, it is generally assumed that the individual neurons that are involved in a distributed representation of an object themselves represent features or properties of the object in question — see, for example, Ref. [29, 92] and references therein. Such representations were called “iconic” by Wickelgren [93, 94].

But another possibility is the use of what Wickelgren [93, 94] called “chunk assemblies”. Each of these is a relatively small set of cells, the activation of which codes the presence or processing of the associated concept. The individual cells in a chunk assembly do not themselves code features corresponding to the concept. Quantitatively, coding using chunk assemblies is characterized by the number of cells in each assembly and by the number of chunk assemblies in which each cell participates. Local coding is the limiting case in which each cell participates in the assembly for exactly one concept, as opposed to participating in merely a very small fraction of the concepts stored in a system. We can regard our work as a step in determining quantitative properties of chunk assemblies. When only a small number of stimuli are used, a cell in a chunk assembly will behave quite similarly to a cell providing a local representation.

We now work out a relation between the sparsity of cell responses, and some of the coding properties. The properties of interest are the total number of concepts coded, the number of concepts that each cell codes (i.e., the number of chunk assemblies it participates in), and the number of concepts that are simultaneously active. Of course neither of the last two numbers needs to be fixed, but it will be useful to treat them as single or representative numbers to get an idea of the relation to sparsity. If only one concept were active at a time, and if the coding were local (so there is only one concept per cell), then the sparsity would be
With the possibilities of multiple concepts being simultaneously active and of more than one concept being coded per cell, we get instead:

$$\text{sparsity} = \frac{(\# \text{ concepts activated})(\# \text{ concepts per cell})}{\text{Total \# concepts}},$$  \hspace{1cm} (4.1)

at least in some average sense, given that both the number of concepts activated and the number stored by the cell are small compared with total number of concepts stored overall. The simplest ideas about local/grandmother-cell representations would assign unity to the number of simultaneously active concepts. In this discussion, we should use the word “concept” very broadly, such that it includes, for example, individual episodic memories.

So how many concepts can a typical human being recognize? An estimate of 10,000 to 30,000 has been quoted by Waydo et al. [78] as the number of objects that a typical adult recognizes, on the basis of work by Biederman [95]. But this is surely a substantial underestimate of the number of concepts that are coded in a human brain. First, the number of words in a language that are known to an adult is in the tens of thousands, and the number of concepts is surely substantially higher. Second, it is known that humans can remember thousands of pictures [96, 97] presented during a single day. The total number of memories created over a lifetime must be orders of magnitude larger. Even allowing that on a long time scale many of these will be forgotten, this indicates that the number of concepts/objects represented can easily be in the hundreds of thousands or millions.

For illustration, assume that the number of concepts remembered is $10^6$. Then a measured sparsity of around $10^{-3}$, as we found for the means of the sparsity distributions for the successful fits from our analysis, implies from (4.1) that

$$\ (\# \text{ concepts activated}) \times (\# \text{ concepts per cell}) \simeq 1000. \hspace{1cm} (4.2)$$

This equation places limits on how close these cells are to grandmother cells, in which case $(\# \text{ concepts per cell}) = 1$. This would require $(\# \text{ concepts activated}) = 1000$ in order to remain consistent with the data analyzed. In other words, each stimulus must simultaneously activate 1000 associated concepts as downstream consequences, which seems unreasonably high. Given the number of concepts we should expect to
be represented in the human brain even our small measured sparsity appears to be much too large to be consistent with a local (grandmother) representation. We leave further analysis to the future, but use this estimate as a suggestion about the quantitative properties of concept coding.

4.1.3 Silent Cells

A second issue [29,80–83] is that there may are likely many silent cells, i.e., cells that gave no detectable spikes at all in the measurements. Silent cells are to be contrasted with the many non-responsive cells that were included and were important in our analysis; non-responsive cells did give detected spikes, but never enough to count as an above-threshold response. Waydo et al. [78] argue that this is potentially a very large effect. They say that “as many as 120–140 neurons are within the sampling region of a tetrode in the CA1 region of the hippocampus”, but only “1–5 units per electrode” were identified.

To quantify the effects of silent cells, let $K$ be the ratio of the total number of cells to the number of detected cells. The suggestion [78,82,92] is that $K$ could be as much as an order of magnitude. The effect of silent cells on our measurement of the sparsity $\alpha_D$ of the higher sparsity population would be minor, because cells in this population typically give at least one above-threshold response and are therefore detected. But their fraction in the total neuronal population, $f_D$ must be decreased by a factor $K$.

For the ultra-sparse population, the sparsity $\alpha_{US}$ is decreased relative to our fits by a factor approximately equal to $K$ (while the population abundance gets closer to 100%). This is simply because $\alpha_{US}$ is, to a first approximation, the ratio of $N_1$ to $N_0 + N_1$. $N_1$ is fixed by data, but $N_0 + N_1$ is increased by a factor of about $K$ to allow for silent cells.

For the beta model, the silent cell factor would bring the sparsity distributions even closer to a divergence at $\alpha = 0$, i.e. the parameter $a$ would be brought even closer to zero. I tested the beta model (with double units) to the hippocampus data assuming $K = 10$. The resulting maximum likelihood estimates of the parameters were $a = 0.007$ and $b = 55$, compared with $a = 0.1$ and $b = 70$ (Table 3.4). The resulting $\chi^2$ value was well within the acceptable range ($\chi^2 = 1.2$ for the first five data points). Taking into consideration the silent cells that should populate the
Figure 4.1. Comparison of the two-population model (blue dots) from Eq (2.12) and the beta model (purple squares) from Eq (2.14) displayed on a log plot. Parameter values are taken from the best fits presented in Chapter 3. Predictions for 100 stimuli are shown in (a), and predictions for 1000 stimuli are shown in (b).

\[ n = 0 \text{ bin, the beta model predicts a sparsity distribution that very nearly achieves scale-free behavior.} \]

The effect of large numbers of silent cells is to substantially strengthen our conclusions, even if some of the numbers are less certain. Of course, if at sometime in the future, a more precise understanding of electrode properties were obtained, then a useful estimate of the silent cell factor \( K \) could be obtained. After that our numerical results could be adjusted accordingly. See [98,99] for recent developments.

4.2 Multiple Models Fit Identical Data

4.2.1 two populations

The two-population model predicts two very different populations of cells in the areas examined. It is possible that these populations are anatomically or functionally different, as with the different kinds of cells in the HVC and RA areas of zebra finches [29,84]. But this is not a necessary deduction. The presence of the different populations might simply represent different semantic properties for the cells’ firing relative to the nature of the particular class of stimuli used. For example, the stimuli might concern musical tunes instead of famous people. Then some cells that had very low sparsity in the one situation could have much higher sparsity in the other situation. Similar phenomena are seen with place field behavior of hippocampal neurons when an animals environment is changed [100]. Since such changing contexts are common experience, it is reasonable to assume that this
situation is typical. The different populations correspond to different semantic domains, and the chosen stimuli sample these domains.

Our inferred populations and sparsities must then be treated as being relative to the general class of stimuli used, e.g., famous individuals, landmarks, animals or objects in the case of the data from [50] that we analyzed.

Now the brain regions probed, e.g., the hippocampus, do not appear to be specialized for particular subject areas. So we should expect that our results should be typical for any choice of subject area. This also makes it acceptable that we analyzed data not from just one patient but data pooled over many patients.

In sharp contrast, the beta model does not suggest the partitioning of cells into distinct populations, but rather suggests that the neural properties are smoothly distributed across a range of values.

4.2.2 Beta model vs Two-population model

The beta model, presented in section 2.3.4, is based upon a continuous probability distribution. On the other hand, the two-population model presented in section 2.3.3 assumes the neurons are split into two discrete populations: a sparse population comprising roughly $5\% - 10\%$ of the MTL neurons, with a sparsity on the order of $10^{-2}$; and an ultra-sparse population comprising the remaining $90\% - 95\%$, with a sparsity on the order of $10^{-3}$. The neurons in each population were assumed to have the same sparsity value. The two-population model produced good fits in the Hipp and the EC, but produced poorer fits in the Amy and the PHC. The beta model, on the other hand, is a continuous distribution, and it fits all four regions adequately, including the Hipp and the EC.

Thus, there are two radically different sparsity distributions that are consistent with the single-unit responses from the Hipp and EC. In order to produce good fits to the data, which shows very large $n_0$ and $n_1$ bins compared with the bins at higher $k$, the sparsity distributions at small $\alpha$ are largely determined by these two bins. There is insufficient statistical power to distinguish between different sparsity distributions in this low sparsity regime. Graphically, this is shown in Fig. 4.1 (a), which shows both models producing similar predictions at $S = 100$. At higher sparsities (Fig. 4.1 (b)), these predictions will differ dramatically.

To illustrate the connection between the beta model and the two-population
Figure 4.2. Best fits to data recorded in four regions of the human MTL assuming all recorded units consist of a single neuron. The red circles indicate the experimental values from Table 3.2 and blue dots connected by lines indicate the best-fit predictions for the expectation values of $n_k$ with double-neuron units included predicted by the beta-binomial model. The dotted lines indicate best fits for a pure binomial model i.e. assuming all cells have the same sparsity. Note the log scale on the y-axis.

model, the shaded region seen in figure 3.6, roughly corresponds to the sparse population (population labeled “D” in section 2.3.3), while the neurons in the unshaded region are analogous to the ultra-sparse population (labeled “US” in section 2.3.3). In the two-population model, the distribution would have a Dirac-delta function located in the shaded region and another located in the unshaded region, representing the two populations.

Consequently, the exact form of the beta distribution should not be taken too literally, as there are likely to be other continuous distributions that are consistent with the data. However, one would still expect similar overall behaviors regardless of which distribution is chosen. The advantage of the beta distribution is that it yields compact analytical results in the likelihood analysis for both the single unit model and the multi-unit model.
4.2.3 Other possible models

As shown in the previous section, the limited number of stimuli make it difficult to distinguish between radically different models, in this case the two-population model and the beta model. It is reasonable to assume that other models could also be used to fit the data presented in [50].

One possibility has already been discussed, is the inclusion of more populations beyond the two considered in section 2.3.3. Another possibility, closely related to the beta-binomial distribution, is the gamma-poisson distribution, which is sometimes called the negative distribution. While the beta-binomial distribution is generated by a binomial distribution, whose probability parameter has been sampled from a beta distribution, the gamma-poisson distribution is generated by a poisson variable whose rate parameter is sampled from a gamma distribution. Thus, the negative binomial is a limiting case of the beta-binomial distribution, much in the same way the poisson is a limiting case of a binomial distribution. This is shown in Appendix B. Thus, like the beta-binomial distribution, it is also skewed and overdispersed.

4.2.4 Speculation on the decoding advantages of a near-power law sparsity distribution

One possible benefit of a near-power law sparsity observed in the beta model (see Fig. 3.6) is that such a sparsity distribution is able to sustain a significant proportion of cells that respond to only one stimulus, even as the set of stimuli grows very large. To show this, $\epsilon_1$ vs $S$ is plotted in Fig. 4.3 for both the beta model and the two-population model using the parameters of the maximum likelihood fits in the hippocampus.

Recalling the benefits of sparse coding described in section 1.1.1, at the highest stages of sensory processing, we would expect an easily-decoded representation of features. Grandmother codes, in which each neuron responds to a single concept, are hypothetically the easiest to decode. Above in Sec. 4.1.2, the number of concepts a human being can reliably recall over a lifetime was estimated to be on the order of $10^6$. In a hypothetical experiment, in which a substantial proportion of the total memorized concepts were presented to the patient, the beta model predicts that $\sim 1/40$ cells would respond to only a single concept, i.e. exhibit characteristics
Figure 4.3. Comparison of the $k = 1$ probabilities between the two-population model (purple curve with a large peak) and the beta model (blue curve). The beta model predicts a significant proportion of cells that respond to only one stimulus, even when a huge number is shown. The two-population model, on the other hand, has a disappearing $k = 1$ bin at large $S$.

of grandmother or gnostic cells. On the other hand, the two population model predicts an utterly negligible ($10^{-43}$) fraction of cells that would possess gnostic cell response characteristics.

Even neglecting any considerations of grandmother cells, the beta model maintains sparse, easily decodable proportions of neural responses as the size of the stimulus set to be represented expands. In other words, it produces a representation with a scalable ease of readout.

4.3 Beta Distributions arising from preferential attachment processes

The success of the beta model warrants consideration with respect to what processes might be responsible for producing this distribution. As mentioned in Sec 1.2.2, Rich et al. [69], found an overdispersed distribution of place fields per place cell in the the rat CA1.

Specifically, they found that the recruitment of place fields among the cells of CA1 obeys a skewed gamma-poisson process, in which each cell acquires place fields according to a poisson process, but the rate parameter for each cell is sampled from a gamma distribution. This results in a poisson distribution of place fields per cell mixed by a gamma distribution.
In this dissertation, (and in [74]), the number of concept fields of MTL neurons are shown above to be distributed by a binomial distribution mixed with a beta distribution. The gamma-poisson distribution and beta-binomial distributions are closely related, with the gamma-poisson being a limiting case of the beta-binomial. For details, see Appendix B.

Place cells of the rat and concept cells in humans share many characteristics [54], and observing nearly identical distributions of activity in both populations suggests that the place cell code and the concept cell code likely arise from similar underlying processes. In other words, the results suggest that the recruitment of place fields by the place cells of the rat hippocampus and the recruitment of “concept fields” by the concept cells of the human MTL are two manifestations of the same neural mechanism, despite coding for spatial information vs conceptual information.

Furthermore, finding nearly identical distributions in different species across cells with strikingly different receptive fields lends more credence to the idea that skewed distributions in general play a fundamental role in neural functioning [70,71].

Rich et al. [69] suggest two explanations for why the place cells might recruit fields at a skewed rate:

1. The skewed field propensities (the rates at which place cells acquire new fields) arise from intrinsic cellular or network differences, e.g. non-uniform connectivity patterns, membrane excitabilities, etc. In this scenario, the field propensities are relatively static over time as the organism navigates the environment.

2. The place fields form in cells according to some preferential attachment or cumulative advantage mechanism. In this scenario, the cells start out with uniform field propensities, but as a cell acquires place fields, its propensity increases, making it more likely to acquire future fields.

They then argued that intrinsic cellular differences as the more likely mechanism, pointing out that the firing rate of the place cells during slow-wave sleep of the rat correlates (albeit weakly) with the number of place fields the cell picked up during navigation.

I would like to suggest that the two possibilities above, for producing a skewed distribution of field propensities, are not mutually exclusive. It is possible that both mechanisms might be at play. In the following discussion, I will interpret
the observed near power-law beta distributions in terms of preferential attachment processes.

4.3.1 Rich get Richer schemes

Beta distributions with near power law behavior arise as limiting distributions in numerous “rich-get-richer” schemes. For example, preferential attachment processes on growing networks yield beta distributions, where nodes with a high degree are more likely to receive edges from newly added nodes [85, 86]. Networks grown in this fashion have been extensively studied, and they possess many favorable properties, such as scale-free node degrees and small network diameters.

Closely related are Polya urn schemes [87], in which balls of different colors are sampled from an urn with super-replacement. Super-replacement is the process of adding additional balls to the urn of the color that was just sampled, so that whenever a color is pulled from urn, the proportion of balls of that color in the urn increase. Networks that grow according to preferential node attachment and Polya urn models both yield beta distributions as \( t \to \infty \).

4.3.2 Bi-layered network model of MTL coding

In this section, we consider the bi-layered network model studied in Peruani et al. [101]. In their model, the bottom layer of the network contains \( N \) nodes which remain fixed in number, while the top layer consists of \( t \) nodes that are added one at a time, starting from \( t = 0 \). As nodes are added, they connect with nodes in the bottom layer, with a higher probability of attaching to nodes with a large degree.

We interpret the fixed bottom layer as the binary neurons of the MTL responsible for concept (or place) representation, and we interpret the top layer as salient concepts or features that are output by the top layers of sensory processing and input into the MTL. The addition of a node in the top layer indicates a new concept that is to be represented, e.g. an unfamiliar person to whom you have just been introduced. An edge between concept \( i \) in the top layer and neuron \( j \) in the bottom layer indicates that activity in neuron \( j \) is evoked by \( i \) (i.e. neuron \( j \) activates whenever stimulus containing \( i \) is presented to the organism). The attachment procedure for new nodes represents the complex neurological processes by which new environmental features are coded onto the neural substrate of the human MTL,
i.e. the formation of a new place field or concept field. The sparsity of neuron $j$ (the fraction of stimuli that trigger activity in the cell), $\alpha_j$, is given by

$$\alpha_j = \frac{k_j^t}{t}$$

(4.3)

where $k_j^t$ is the node degree of neuron $j$ after $t$ stimuli have been added in the top layer.

### 4.3.3 Growing the network

Following Peruani, et al. [101], the network is grown by adding a node to the top layer at each time step, $t$, and then attaching it to $\mu$ nodes in the bottom layer. In our model, $\mu$ is the code word length of each stimulus and is assumed to be fixed. We assume $\mu \ll N$. As each of the $\mu$ edges is added, the probability $A \left( k_j^t \right)$ that it attaches to neuron $j$ with node degree $k_j^t$ is defined by

$$A \left( k_j^t \right) = \frac{1}{C(t)} \left( \gamma k_j^t + 1 \right)$$

(4.4)

where $\gamma$ is a parameter that determines the influence of node degree on the attachment process and $C(t)$ is a normalization constant. For $\gamma = 0$, the edges are attached at random to the neurons with equal probability, $A \left( k_j^t \right) = \frac{1}{N}$ for each $j$.

In the case where $\mu \ll N$, the probability that neuron $j$ connects to the stimulus added at time $t$ after all $\mu$ edges have been connected is approximately

$$A_\mu \left( k_j^t \right) = \frac{1}{C(t)} \mu \left( \gamma k_j^t + 1 \right)$$

(4.5)

When $\gamma = 0$, a newly added stimulus attaches to neuron $j$ with probability $A_\mu \left( k_j^t \right) = \frac{\mu}{N}$.

The network is grown starting from $t = 0$, with no stimuli in the top layer and all node degrees in the bottom layer equal zero. Eq. (4.5) defines the attachment process as each stimulus is added. We are interested in the sparsity distribution, $D \left( \frac{k}{t} \right)$ of neurons in the bottom layer after a large number of stimuli have been learned. Peruani et al. [101] showed that for $\gamma > 0$, (derivation in Appendix C) in
the limit of large \( t \),

\[
D \left( \frac{k}{t} \right) = \frac{1}{B(r,s)} \left( \frac{k}{t} \right)^{r-1} \left( 1 - \frac{k}{t} \right)^{s-1}
\]  
(4.6)

where \( r = \frac{1}{\gamma} \) and \( s = \frac{N}{\gamma\mu} - \frac{1}{\gamma} \).

In the case of purely random attachment (\( \gamma = 0 \)), the distribution \( D \left( \frac{k}{t} \right) \to \delta \left( \frac{k}{t} - \frac{\mu}{N} \right) \) as \( t \to \infty \), i.e. the neuronal sparsity is the same for all neurons. Purely random attachment is inconsistent with the skewed Mormann data we fit and the overdispersed place field data from Rich et al.

If we treat the data analyzed above as a random sample of \( N \) neurons from the bottom layer and \( S \) stimuli from the top layer, then we can match the beta parameters \( r \) and \( s \) with the beta fits reported in the previous section. This gives an estimate of roughly \( 10 \leq \gamma \leq 20 \) and \( \frac{1}{600} \leq \frac{\mu}{N} \leq \frac{1}{250} \). These values for \( \gamma \) suggest from Eq. (4.5) that preferential attachment plays a large role in assigning new stimuli to MTL neurons (assuming that the feature coding follows the bi-layered model presented). In other words, it is indirect evidence that neurons are not randomly assigned to new concepts, but rather new concepts are more likely to be coded onto neurons that have previously been assigned to previously learned concepts.
Chapter 5  |  Conclusions

In this dissertation, I presented the statistical approach described in Magyar and Collins [73] and Magyar [74] for analyzing the single-unit recordings taken from the human MTL presented in Mormann et al. [50]. The approach treats the binary units and the presented stimuli as statistical samples and uses the recorded responses to estimate the distribution of neural sparsity across the cells of the human MTL. The response probability is termed the neuronal sparsity and is the central concept of this model that sets it apart from other analyses of this sort of data, which often apply descriptive rather than inferential statistics.

Applying this approach to test multiple distributions reveals a great deal about the quantitative aspects of the neural code in the MTL. The resulting fits indicate that the code is extremely sparse and heavily skewed, with a wide range of response probabilities amongst the cells. The nature of the fits allow us to estimate how close the concept cells are to grandmother cells, and supplies predictions for what the neural responses might be when large numbers of stimuli are presented. The issues of silent cells and multi-units were taken into careful consideration, which both serve to strengthen the case for sparse coding in the human MTL.

Specifically, the two-population model and the beta model were shown to produce good fits in the hippocampus and entorhinal cortex, while the beta model also produced good fits in the amygdala and the parahippocampal cortex. The beta model fits have a near-power law divergence at low sparsity, suggesting preferential attachment as a possible generating mechanism. Furthermore, the beta model allows for a significant proportion of cells to code for a small number of concepts, even for stimulus sets large enough to challenge the capacity of the brain, suggesting that a beta code might scale in such a way as to maintain ease of decodability.
Also, the sparsity distributions of the beta model show remarkable similarities to the statistics of place field propensities in the rat CA1, suggesting deep connections between the dynamics underlying the generation of the place field code and the concept code in humans.

Although the model was applied to one experiment in the human MTL, it could in principle be applied to any data set in which the recorded units are independent and exhibit binary all-or-nothing responses to a sequence of stimuli. In particular, applying the methods to place cell statistics in the rat hippocampus would be an excellent avenue for future work.

The MTL, however, is particularly interesting to study. The MTL receives the output of sensory processing across several modalities. At this stage, the cells are responding to complex, behaviorally relevant features of the environment, and may play a central role in perception and cognition. Understanding the statistics of the responses of these cells is an early step towards developing and constraining a full theory of how sensory processing is organized.
Appendix A
Derivation of Poisson approximation to likelihood function

A.1 Introduction

Here we provide a derivation of the Poisson approximation

$$\mathcal{L} \approx \prod_{k=1}^{S} e^{-N\epsilon_k} \frac{(N\epsilon_k)^{n_k}}{n_k!}, \quad (A.1)$$

to the likelihood function

$$\mathcal{L}({\alpha_i, f_i} \mid \text{Data}) = \frac{N!}{\prod_{k=0}^{S} n_k!} \prod_{k=0}^{S} \epsilon_k^{n_k}, \quad (A.2)$$
given the normalization conditions

$$\sum_{k=0}^{S} \epsilon_k = 1, \quad \sum_{k=0}^{S} n_k = N. \quad (A.3)$$

The derivation applies when the values of $\epsilon_k$ for non-zero $k$ are small, more precisely, when $\sum_{k=1}^{S} \epsilon_k \ll 1$.

We start from Eq. (2.16) by factoring out the $k = 0$ factors from the product, and then using Eq. (A.3) to write $n_0$ and $\epsilon_0$ in terms of the corresponding quantities for $k \geq 1$:

$$\mathcal{L}({\alpha_i, f_i}) = \frac{N!}{(N - n_{\geq 1})!} (1 - \epsilon_{\geq 1})^{N-n_{\geq 1}} \prod_{k=1}^{S} \frac{\epsilon_k^{n_k}}{n_k!}, \quad (A.4)$$
where \(n \geq 1 = \sum_{k=1}^{S} n_k\), and \(\epsilon \geq 1 = \sum_{k=1}^{S} \epsilon_k\). By taking the logarithm of both sides of Eq. (A.4) we get:

\[
\ln \mathcal{L} = \ln N! - \ln[(N - n \geq 1)!] + (N - n \geq 1) \ln(1 - \epsilon \geq 1) + \ln \prod_{k=1}^{S} \frac{\epsilon_k^{n_k}}{n_k!},
\]

(A.5)

For large \(N\), we can use Stirling’s approximation to \(O(1/N)\):

\[
\ln N! = N \ln N - N + \frac{1}{2} \ln(2\pi N) + O\left(\frac{1}{N}\right).
\]

(A.6)

Applying this formula to the first two terms in (A.5) yields

\[
\ln \mathcal{L} = N \ln N + (N - n \geq 1) \ln \left(\frac{1 - \epsilon \geq 1}{N - n \geq 1}\right) - n \geq 1 + \ln \prod_{k=1}^{S} \frac{\epsilon_k^{n_k}}{n_k!} + O\left(\frac{1}{N}\right)
\]

\[
= n \geq 1 \ln N + (N - n \geq 1) \ln \left(1 - \frac{\epsilon \geq 1 - n \geq 1/N}{1 - n \geq 1/N}\right) - n \geq 1 + \ln \prod_{k=1}^{S} \frac{\epsilon_k^{n_k}}{n_k!} + O\left(\frac{1}{N}\right).
\]

(A.7)

The approximation worsens beyond the order \(1/N\) error estimate if \(n \geq 1\) gets close to \(N\). But since the \(\epsilon_k\) are small, this situation is of very low probability. Henceforth the error estimates will apply not too far from the peak of the likelihood distribution.

The first term can be combined with the product term

\[
n \geq 1 \ln N + \ln \prod_{k=1}^{S} \frac{\epsilon_k^{n_k}}{n_k!} = \ln \left[N^{n \geq 1} \prod_{k=1}^{S} \frac{\epsilon_k^{n_k}}{n_k!}\right] = \ln \prod_{k=1}^{S} \frac{(N\epsilon_k)^{n_k}}{n_k!}.
\]

(A.8)

We can simplify the remaining logarithm in Eq. (A.7)

\[
(N - n \geq 1) \ln \left(1 - \frac{\epsilon \geq 1 - n \geq 1/N}{1 - n \geq 1/N}\right) = (N - n \geq 1) \left(\frac{n \geq 1/N - \epsilon \geq 1}{1 - n \geq 1/N}\right) + O\left(\frac{(\epsilon \geq 1 - n \geq 1/N)^2}{1 - n \geq 1/N}\right).
\]

(A.9)

For a multinomial distribution,

\[
\epsilon \geq 1 - n \geq 1/N = \sum_{k=1}^{S} (\epsilon_k - n_k/N) = \sum_{k=1}^{S} O\left(\sqrt{\epsilon_k/N}\right),
\]

(A.10)

where we have used the standard deviation of the distribution to estimate the typical value of \(|\epsilon_k - n_k/N|\).
Thus we can write:

\[
O \left( \frac{\epsilon_{\geq 1} - n_{\geq 1}/N}{1 - n_{\geq 1}/N} \right)^2 = O \left[ \frac{1}{N} \left( \sum_{k=1}^{S} \sqrt{\epsilon_k} \right)^2 \right] = O(\epsilon_{\geq 1}/N). \tag{A.11}
\]

Thus, Eq. (A.9) can be simplified to:

\[
(N - n_{\geq 1}) \ln \left( \frac{1 - \epsilon_{\geq 1} - n_{\geq 1}/N}{1 - n_{\geq 1}/N} \right) = n_{\geq 1} - N\epsilon_{\geq 1} + O(\epsilon_{\geq 1}) + O(1/N). \tag{A.12}
\]

Substituting Eqs. (A.8) and (A.12) into Eq. (A.7) yields

\[
\ln L = -N\epsilon_{\geq 1} + \ln \prod_{k=1}^{S} \frac{(N\epsilon_k)^{n_k}}{n_k!} + O(\epsilon_{\geq 1}) + O(1/N). \tag{A.13}
\]

Therefore the likelihood function in our approximation is given by a product of Poisson distributions:

\[
L \approx \prod_{k=1}^{S} e^{-N\epsilon_k} \frac{(N\epsilon_k)^{n_k}}{n_k!}, \tag{A.14}
\]

which is Eq. (2.20).
Appendix B

Gamma-Poisson distribution as a limiting case of the beta-binomial distribution

In this appendix, I show that the gamma-poisson distribution is a limiting case of the beta-binomial distribution. I show it here because a convenient reference showing this result could not be located.

The beta-binomial distribution is a mixture distribution of a binomial distribution, with parameters $S$ and $0 < \alpha < 1$, where the probability parameter $\alpha$ is sampled from a beta distribution with parameters $a > 0$ and $b > 0$. The PMF for the binomial distribution and the PDF for the beta distributions are given respectively:

$$P(K = k) = \binom{S}{k} \alpha^k (1 - \alpha)^{S-k} \quad \text{(B.1)}$$

and

$$D(\alpha) = \frac{\alpha^{a-1} (1 - \alpha)^{b-1}}{B(a,b)} \quad \text{(B.2)}$$

respectively.

Thus, the PMF of the beta-binomial distribution with parameters $S$, $a$, and $b$ is given by:

$$P(K = k) = \frac{\binom{S}{k}}{B(a,b)} \int_0^1 d\alpha \alpha^k (1 - \alpha)^{S-k} \alpha^{a-1} (1 - \alpha)^{b-1}$$

$$= \frac{\binom{S}{k} B(a+k, b+S-k)}{B(a,b)}. \quad \text{(B.3)}$$
The gamma-poisson distribution, more commonly called the negative binomial distribution, is a Poisson distribution with parameter $\lambda > 0$ where $\lambda$ is sampled from a gamma distribution with parameters $r > 0$ and $\beta > 0$. The PMF of the Poisson distribution and the PDF of the gamma distribution is given:

$$
P(K = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{(B.4)}$$

and

$$
g(\lambda) = \frac{r^\beta}{\Gamma(\beta)} \lambda^{\beta-1} e^{-r\lambda} \quad \text{(B.5)}$$

and the gamma-poisson distribution with parameters $r$ and $\beta$ is then given by

$$
P(K = k) = \frac{1}{k!} \frac{r^\beta}{\Gamma(\beta)} \int_0^\infty d\lambda \frac{\lambda^{k+\beta-1} e^{-\lambda(1+r)}}{\lambda^{r+1}}$$

$$
= \frac{1}{k!} \frac{r^\beta}{\Gamma(\beta)} \int_0^\infty dx \frac{x^{k+\beta-1}}{r+1} \frac{1}{(r+1)^{k+\beta}} e^{-x}$$

$$
= \frac{1}{k!} \frac{r^\beta}{\Gamma(\beta)} \frac{\Gamma(k+\beta)}{(r+1)^{k+\beta}} \quad \text{(B.6)}$$

To show that B.3 yields B.6 as a limiting case, we begin by observing that the Poisson distribution is a limiting case of the binomial distribution as $S \to \infty$ while holding the mean, $\langle k \rangle = S\alpha$ constant. The Poisson parameter $\lambda$ is the expected response, i.e. $\lambda = \langle k \rangle$.

Similarly, the gamma-Poisson distribution is the limiting case of the beta-binomial distribution as $S \to \infty$ while ensuring the mean $\langle k \rangle = S\langle \alpha \rangle$ for the beta-binomial distribution is finite. From Eq. (B.2), we see that

$$
\langle \alpha \rangle = \frac{1}{B(a,b)} \int_0^1 \alpha^a (1-\alpha)^{b-1} = \frac{B(a+1,b)}{B(a,b)} = \frac{a}{a+b} \quad \text{(B.7)}
$$

So,

$$
\langle k \rangle = S\langle \alpha \rangle = \frac{Sa}{a+b} \quad \text{(B.8)}
$$

One way to ensure $\langle k \rangle$ is finite as $S \to \infty$, is for the parameter $b$ to approach infinity as a linear function of $S$ by setting $b = rS$. The constant $r > 0$ is arbitrary.
It will be shown that that $r$ matches the parameter of the gamma distribution above when the limit is taken.

Before taking the limit by setting $b = rS$ in Eq. (B.3), it is useful to write the beta functions and binomial coefficient in terms of gamma functions using the identities

$$\binom{S}{k} = \frac{\Gamma(S + 1)}{\Gamma(k + 1)\Gamma(S - k + 1)}$$  \hspace{1cm} (B.9)$$

and

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}$$  \hspace{1cm} (B.10)$$

Using these identities, we get write Eq. (B.3) as

$$P(K = k) = \frac{\Gamma(S + 1)}{\Gamma(k + 1)\Gamma(S - k + 1)} \times \frac{\Gamma(a + k)\Gamma(S - k + b)}{\Gamma(a + b + S)} \times \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}$$

$$= \frac{\Gamma(S + 1)}{\Gamma(k + 1)\Gamma(S - k + 1)} \times \frac{\Gamma(a + k)\Gamma(S(1 + r) - k)}{\Gamma(a + S(1 + r))} \times \frac{\Gamma(a + rS)}{\Gamma(a)\Gamma(rS)}$$  \hspace{1cm} (B.11)$$

Now, to take the limit as $S \to \infty$, we use Stirling’s approximation applied to ratios of gamma functions:

$$\frac{\Gamma(x + \gamma)}{\Gamma(x + \omega)} \approx x^{\gamma - \omega}$$  \hspace{1cm} (B.12)$$

for large $x$. We get:

$$P(K = k) = \frac{S^k}{\Gamma(k + 1)} \times \frac{\Gamma(a + k)}{[S(1 + r)]^{a+k}} \times \frac{(rS)^a}{\Gamma(a)}$$

$$= \frac{1}{k!} \times \frac{\Gamma(a + k)}{(1 + r)^{a+k}} \times \frac{r^a}{\Gamma(a)}$$  \hspace{1cm} (B.13)$$

Comparing Eqs (B.13) and (B.6), we see that they match, and that $a = \beta$. This concludes the derivation.
Appendix C
Beta Distribution as the limiting distribution of preferential attachment

Here, the mathematical details surrounding the result reported in Peruani et al. [101] will be fleshed out. Specifically, it will be shown that a bi-layered network with one fixed layer and one growing layer tends towards a beta distribution of node degrees, $k_j^t$, in the bottom layer under preferential attachment. The subscript $j$ indexes the nodes (neurons) of the bottom layer.

We proceed in two steps: first we derive the node degree distribution of the bottom layer after $t$ nodes (fields) have been added to the top layer, then we take the limit as $t \to \infty$ and show that the distribution of sparsity $\alpha = \frac{k_j^t}{t}$ across the bottom layer nodes approaches a beta distribution.

C.1 The node degree distribution

To begin, we define the attachment process. For the remainder, I refer to the $N$ nodes of the fixed layer as neurons, and nodes of the growing layer as fields (in analogy to concept fields or place fields—the things to be represented that are input into the MTL by the top layer of sensory processing). At $t = 0$, there are no fields added to the network and $k_j^t = 0$ for all neurons $j$. At the introduction of each field, $\mu$ edges are connected to neurons. Define $A_\mu(k_j; t)$ as the conditional probability that node $j$ receives an edge during the addition of the new field at time $t$, given
its node degree, $k^t_j$: 

$$A_{\mu} (k^t_j, t) = \frac{1}{C(t)} \mu \left( \gamma k^t_j + 1 \right)$$  \hspace{1cm} (C.1)$$

where $C(t) = \gamma \mu t + N$ is a normalization factor (the sum of Eq C.1 over all neurons $j$ should equal $\mu$) and $\gamma \geq 0$ is the parameter that governs the influence of node degree on the attachment probability. Let $p_{k,j,t}$ be the total probability that neuron $j$ has node degree $k$ after $t$ fields have been formed. To derive $p_{k,t}$ (note the subscript $j$ has been dropped for ease of notation), we need to calculate the probability that $k$ attachment events connect with neuron $j$ out of $t$ overall.

First, let $i_1, i_2, \ldots, i_k$ be the time indices for the attachment events. In other words, at time $t = i_1$, node $j$ receives its first edge; at time $t = i_2$ it receives its second, and so on. We first calculate the probability of this particular sequence of events using Eq (C.1):

$$P\left(\text{sequence}\right) = \left[1 - A_{\mu} (0, 0)\right] \times \left[1 - A_{\mu} (0, 1)\right] \times \ldots \times \left[1 - A_{\mu} (0, i_1 - 2)\right] A_{\mu} (0, i_1 - 1)$$

$$\times \left[1 - A_{\mu} (1, i_1)\right] \times \left[1 - A_{\mu} (1, i_1 + 1)\right] \times \ldots \times \left[1 - A_{\mu} (1, i_2 - 2)\right] A_{\mu} (1, i_2 - 1)$$

$$\times \ldots$$

$$\times \left[1 - A_{\mu} (k - 1, i_{k-1})\right] \times \ldots \times \left[1 - A_{\mu} (k - 1, i_k - 2)\right] A_{\mu} (k - 1, i_k - 1)$$

$$\times \left[1 - A_{\mu} (k, i_k)\right] \times \left[1 - A_{\mu} (k, i_k + 1)\right] \times \ldots \times \left[1 - A_{\mu} (k, t - 2)\right] \times \left[1 - A_{\mu} (k, t - 1)\right]$$  \hspace{1cm} (C.2)$$

The first row of factors are the attachments events up to the first successful event that attaches to neuron $j$, and so on for each following row. Simplifying, we get:

$$P\left(\text{sequence}\right) = \prod_{s=0}^{i_1-2} \left[1 - A_{\mu} (0, s)\right] \times \prod_{s=i_1}^{i_2-2} \left[1 - A_{\mu} (1, s)\right] \times \ldots \times \prod_{s=i_{k-1}}^{i_k-2} \left[1 - A_{\mu} (k - 1, s)\right]$$

$$\times \prod_{s=i_k}^{t-1} \left[1 - A_{\mu} (k, s)\right] \times \prod_{v=0}^{k-1} \left[A_{\mu} (v, i_{v+1} - 1)\right]$$  \hspace{1cm} (C.3)$$

To simplify further we define some useful functions, $a(k)$ and $b(t - k)$

$$A_{\mu} (k^t_j, t) = \frac{1}{C(t)} \mu \left( \gamma k^t_j + 1 \right)$$

$$\equiv \frac{a(k)}{C(t)}$$  \hspace{1cm} (C.4)$$
\[ 1 - A_\mu(k, j, t) = \frac{C(t) - \mu \left( \gamma k_j^t + 1 \right)}{C(t)} = \frac{\gamma \mu (t - k - \gamma^{-1}) + N}{C(t)} \equiv \frac{b(t - k)}{C(t)} \]  \hspace{1cm} (C.5)

Writing Eq (C.3) in terms of \( a(k) \), \( b(t - k) \), and \( C(t) \):

\[
P(\text{sequence}) = \prod_{s=0}^{i_1 - 2} \frac{b(s)}{C(s)} \times \prod_{s=i_1}^{i_2 - 2} \frac{b(s - 1)}{C(s)} \times \cdots \times \prod_{s=i_{k-1}}^{i_k - 2} \frac{b(s - k + 1)}{C(s)} \]

\[
\times \prod_{s=i_k}^{t-1} \frac{b(s - k)}{C(s)} \times \prod_{v=0}^{k-1} \frac{a(v)}{C(i_v - 1)}
\]

\[
= \frac{\prod_{s=0}^{k-1} a(s) \times \prod_{s'=0}^{t-k-1} b(s')}{\prod_{s=0}^{t-1} C'(s)}
\]  \hspace{1cm} (C.6)

Note that Eq (C.6) is independent of the time indices for the attachment events, \( i_1, i_2, \ldots, i_k \). Thus, the probability is the same for all sequences of attachments that result in \( k \) edges attaching to node \( j \). There are \( \binom{i}{k} \) such sequences, so the probability of node \( j \) having \( k \) degree after \( t \) attachments is given by;

\[
p_{k,t} = \binom{t}{k} \frac{\prod_{s=0}^{k-1} a(s) \times \prod_{s'=0}^{t-k-1} b(s')}{\prod_{s=0}^{t-1} C'(s)}
\]

\[
= \binom{t}{k} \frac{\prod_{s=0}^{k-1} \mu (\gamma s + 1) \times \prod_{s'=0}^{t-k-1} \gamma \mu (s' - \gamma^{-1} + N)}{\prod_{s=0}^{t-1} (\gamma \mu s + N)}
\]

\[
= \binom{t}{k} \frac{\prod_{s=0}^{k-1} (s + \gamma^{-1}) \times \prod_{s'=0}^{t-k-1} (s' - \gamma^{-1} + \frac{N}{\gamma \mu})}{\prod_{s=0}^{t-1} \left( s + \frac{N}{\gamma \mu} \right)}
\]  \hspace{1cm} (C.7)

\[
= \binom{t}{k} \frac{\langle \gamma^{-1} \rangle_k \langle \frac{N}{\gamma \mu} - \gamma^{-1} \rangle_{t-k}}{\langle \frac{N}{\gamma \mu} \rangle_{t}}
\]  \hspace{1cm} (C.8)

The final equality, Eq (C.8), is written using the rising factorial, \( \langle x \rangle_n = x(x + 81 \]
1)(x + 2)⋯(x + n − 1). Eq (C.7) corresponds to the result obtained in Peruani et al. [101] (their Eq (11)) derived by different methods. Eq (C.8) is the node degree distribution of the neurons after $t$ attachments. Next, we show that in the limit as $t \to \infty$, the random variable $\frac{k}{t}$ approaches a beta-distributed random variable.

### C.2 Deriving the beta distribution

The goal of this section is to use the node degree distribution of the fixed layer derived in C.8 to show that the sparsity, $\alpha = \frac{k}{t}$ approaches a beta distribution as the top layer grows to infinity.

The challenge is to transition from the discrete node degree distribution to a continuous beta distribution. To do this, we consider the cumulative distribution function for the node degree, $P(K \leq k) = \sum_{s=0}^{k} p_{s,t}$. To introduce the sparsity, we write

$$ P(K \leq t\alpha) = \sum_{s=0}^{[t\alpha]} p_{s,t}. \quad (C.9) $$

To proceed, we write the rising factorials of Eq C.8 in terms of gamma functions using the identity

$$ \langle x \rangle_n = \frac{\Gamma(x + n)}{\Gamma(x)} \quad (C.10) $$

and Eq B.9 we get:

$$ P(K \leq t\alpha) = \sum_{s=0}^{[t\alpha]} \left( \frac{t}{k} \frac{\langle \gamma^{-1} \rangle_k \langle \eta \rangle_{t-k}}{\langle \eta + \gamma^{-1} \rangle_t} \right) $$

$$ = \sum_{s=0}^{[t\alpha]} \left( \frac{t}{k} \frac{\langle \eta \rangle_{t-k}}{\langle \eta + \gamma^{-1} \rangle_t} \right) $$

$$ = \sum_{s=0}^{[t\alpha]} \left\{ \frac{\Gamma(t + 1)}{\Gamma(k + 1)\Gamma(t - k + 1)} \times \frac{\Gamma(\gamma^{-1} + k)}{\Gamma(\gamma)} \right\} $$

$$ \times \frac{\Gamma(\eta + t - k)}{\Gamma(\eta)} \times \frac{\Gamma(\eta + \gamma^{-1})}{\Gamma(\eta + \gamma^{-1} + t)} \right\} \quad (C.11) $$

where $\eta = \frac{N}{\gamma t} - \gamma^{-1}$. By applying Stirling’s approximation to ratios of gamma
functions, Eq (B.12), we get:

\[
P(K \leq t\alpha) = \sum_{s=0}^{[t\alpha]} \frac{\Gamma(\eta + \gamma^{-1})}{\Gamma(\eta)\Gamma(\gamma^{-1})} \times \frac{(t-k)^{\eta-1}k^{\gamma-1}}{t^{\eta+\gamma-1}}
\]

\[
= \frac{\Gamma(\eta + \gamma^{-1})}{\Gamma(\eta)\Gamma(\gamma^{-1})} \sum_{s=0}^{[t\alpha]} \frac{1}{t} (\alpha)^{\gamma-1-1} (1 - \alpha)^{\eta-1}
\]

(C.13)

Finally, by taking the limit \( t \to \infty \), the sum becomes a Riemann sum, and thus becomes an integral:

\[
P(K/t \leq \alpha) = \frac{1}{B(\gamma^{-1}, \eta)} \int_0^\alpha \frac{d\alpha'}{(\alpha')^{\gamma-1}} (1 - \alpha')^{\eta-1}
\]

(C.14)

We recognize this as the CDF of the beta distribution. This concludes the demonstration that the beta function is the limiting case of preferential attachment on a bi-layered network with one fixed layer.
Appendix D | Error Analysis

Suppose that we have estimated the best fit parameters by maximizing likelihood with respect to the parameters \( \{\alpha_i, f_i\} \). Then taking the likelihood function to be approximately Gaussian in the vicinity of its maximum, we can estimate the confidence intervals and correlations of the model parameters \([88]\) from the diagonal terms of the covariance matrix, defined by:

\[
\text{cov}(\theta_i, \theta_j) = - (H^{-1})_{ij},
\]  

(D.1)

where the Hessian matrix is

\[
H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln \mathcal{L}.
\]  

(D.2)

The diagonal terms yield the parameter variances, \( \text{cov}(\theta_i, \theta_i) = \sigma_i^2 \). Correlations between model parameters, \( \rho_{ij} \), can be determined from the off-diagonal elements of Eq. (D.1),

\[
\rho_{ij} = \frac{\text{cov}(\theta_i, \theta_j)}{\sigma_i \sigma_j}.
\]  

(D.3)

Two important properties of a Poisson distribution \( e^{-N \epsilon_k} (N \epsilon_k)^{n_k} / n_k! \) are its mean and standard deviation

\[
\langle N_k \rangle = N \epsilon_k, \quad \text{s.d.}(N_k) = \sqrt{N \epsilon_k}.
\]  

(D.4)

The uncertainties are reported in Table 3.2. Correlations between the three parameters of the full two-population model were calculated using Eq. (D.3), and are shown in Table D.1.
Table D.1. Correlations between parameters of the two-active population model, assum-
ing all recorded units are comprised of a single neuron. The presence of units consisting of
two neurons did not affect the correlations between parameters to within two significant
digits.

<table>
<thead>
<tr>
<th></th>
<th>$\rho_{US,SD}$</th>
<th>$\rho_{US,F}$</th>
<th>$\rho_{SD,F}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hipp</td>
<td>0.46</td>
<td>-0.52</td>
<td>-0.62</td>
</tr>
<tr>
<td>EC</td>
<td>0.34</td>
<td>-0.33</td>
<td>-0.41</td>
</tr>
<tr>
<td>Amy</td>
<td>0.33</td>
<td>-0.31</td>
<td>-0.37</td>
</tr>
<tr>
<td>PHC</td>
<td>0.30</td>
<td>-0.23</td>
<td>-0.19</td>
</tr>
</tbody>
</table>
Bibliography


Vita
Andrew Magyar

_Eduction

• Ph.D Pennsylvania State University (2009-2015)
• B.S. Ohio Northern University (2005-2009)

_Papers

• “Two-population model for MTL neurons: the vast majority are almost silent” A. Magyar, J.C. Collins. Phys. Rev. E. 96 012712
• “Beta distribution of human MTL sparsity: a sparse and skewed code” A. Magyar (in preparation)