1-1-2015

Efficient Computation in the Brain

Xuexin Wei
University of Pennsylvania, weixxpku@gmail.com

Efficient Computation in the Brain

Abstract
It has been long proposed that the brain should perform computation efficiently to increase the fitness of the organism. However, the validity of this prominent hypothesis remains debated. In this thesis, I investigate how this idea of efficient computation can guide us to understand the operational regimes underlying various cognitive functions, in particular perception and spatial cognition. In the first study, I demonstrate that such idea leads to a well-constrained yet powerful model framework for human perceptual behaviors by assuming the system is efficient both in term of encoding and decoding. This framework, when applying to human visual perception, explains many reported perceptual biases, including the repulsive biases away from prior peak, which are counter-intuitive according to the traditional Bayesian view. This framework also offers a principle way to address the common criticisms of Bayesian models in perception, which argue that Bayesian models are lack of constraints. In the second study, I demonstrate that the idea of efficiency, coupled with a few assumptions, allows us to make quantitative predictions on the functional architecture of the grid cell system in rodents. One such prediction is that the spatial scales of grid modules should follow a geometric progression, importantly, with the scaling factor to be close to the square root of transcendental number e ~1.6. Such zero-parameter predictions closely match the data reported in recent neurophysiological experiments. The theory also makes several other predictions, some of which have been confirmed by the data. This study suggests that achieving efficiency computation may also apply to neural circuits involving a high-level cognition, i.e. representation of space. In the third study, I analytically derive a generic connection between mutual information and Fisher information. This clarifies an important theoretical issue which has been misunderstood in previous neural coding literature. Additionally, it provides some powerful signatures of the Efficient coding hypothesis, which could guide future experimental tests. Together, the results presented in this thesis suggest that achieving efficient computation serves as a basic design principle which generalizes across neural systems processing low-level and high-level cognitive functions.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Psychology

First Advisor
Alan A. Stocker

Second Advisor
Vijay Balasubramanian

Keywords
Bayesian inference, computational principles, Efficient coding, grid cells, neural computation, perception

EFFICIENT COMPUTATION IN THE BRAIN

Xue-Xin Wei

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

Alan Stocker
Assistant Professor of Psychology

Co-Supervisor of Dissertation

Vijay Balasubramanian
Professor of Physics

Graduate Group Chairperson

John Trueswell, Professor of Psychology

Dissertation Committee:
David Brainard, Professor of Psychology
Russell Epstein, Professor of Psychology
Eero Simoncelli, Professor of Neural Science, NYU

Dedication

To my parents, Jun-Gang Wei and Xiu-Zhen Wei.

To my wife, Zi-Juan Chen.

Acknowledgments

I would like to thank my advisors, Prof. Alan Stocker and Prof. Vijay Balasubramanian for their encouragement and support over the five-year Ph. D. period. Over time, I come to realize that Alan and Vijay differ in many ways in terms of the scientific style. However, there is one thing in common. That is to look for a simple and principled explanation underlying a set of seemingly diverse and complicated observations. I guess due to this reason, I often feel a sense of beauty and elegance in their science. I have benefited greatly from both of them.

I would like to thank Prof. David Brainard and Prof. Russell Epstein for their support throughout. As the Chair of my thesis committee, David always finds the easiest way to solve the problem every time I have a problem. Russell is always there for my random questions. I also want to thank the faculty members in the System Neuroscience group, including Prof. Josh Gold, Yale Cohen, Nicole Rust, Johannes Burge, Maria Geffen for various feedbacks on the materials presented in this thesis. Particularly, I acknowledge Josh for his valuable comments on Chapter 3 and another manuscript which is not included in this thesis, as well as spending time improving my talks.

I thank members in the CPC lab and the physics of living matter group for many interesting interactions. In particular, I want to thank Jason Prentice, for thinking about grid cells together these years. I am also especially grateful to Matjaz Jogan. He was almost always the first person listening to my stuff after I discovered something. I also want to acknowledge Pedro Ortega for sharing his knowledge about cybernetics and the ubiquitousness of rate-distortion theory. I thank Ann Hermundstad, who has kindly sent me her thesis as a template. I thank Alex Tank, Marcelo Mattar, Toni Saarela, Long Luu, Kristy Simmons, John Briguglio, Kamesh Krishnamurthy, Jan Homann for many fun conversations.

Special thanks to Yu Hu, who has been a good friend of mine since college, and now also a wonderful colleague. To Zhuo Wang, for many conferences we attend together and for the helpful feedback on many aspects of this thesis, particularly Chapter 5. To Taisong Jing, Tong Li, Xingtan Zhang for the friendship since college

we did together in Peking University. I want to thank the folks in Penn Computational Neuroscience journal clubs for every piece discussed together, including but not limited to Jeremy Manning, Yin Li, Marino Pegan, Drew Jaegle.

I want to thank all the people I met in the Cold Spring Harbor summer school in 2014. I had a wonderful experience there. In particular, I want to thank Bruno Olshausen, Greg DeAngelis, Greg Horwitz, Stefan Treue for the discussions on various work presented in this thesis. I want to thank Jonathan Pillow, Ann Churchland, Stephen Palmer, Geoffrey Boynton and others for many great games played together.

I have been very lucky to meet and benefit from many great scientists outside UPenn during these years. The list is long so I shall not list it. Especially, I want to mention Prof. Eero Simoncelli, who always comes to my presentation in conferences. His constructive criticisms and insights have shaped this thesis in various ways. I also thank Eero for being an external member of my dissertation committee and for his careful reading of this thesis.

I am grateful to my family. My wife, Zi-Juan Chen, has always been supportive of me doing science, particularly during those difficult moments for me. Actually, if not for her, I probably would not ending up pursuing a Ph. D in UPenn. A substantial portion of the results presented here were obtained while sitting the train commuting to New Jersey, at the time when she was there. Special thanks to our dog, *Small*, who is now a three-year-old lovely Saint Barnard, for all the joy he brings to us.

Last but not least, I should thank the bench in front of the statue of Benjamin Franklin in Penn, which will always be part of my memory. Most of the results in this thesis was emerged and completed sitting on the bench between spring and fall, while enjoying the beautiful view from there.

ABSTRACT

EFFICIENT COMPUTATION IN THE BRAIN

Xue-Xin Wei

Alan Stocker

Vijay Balasubramanian

It has been long proposed that the brain should perform computation efficiently to increase the fitness of the organism. However, the validity of this prominent hypothesis remains debated. In this thesis, I investigate how this idea of *efficient computation* can guide us to understand the operational regimes underlying various cognitive functions, in particular perception and spatial cognition. In the first study, I demonstrate that such idea leads to a well-constrained yet powerful model framework for human perceptual behaviors by assuming the system is *efficient* both in term of *encoding* and *decoding*. This framework, when applying to human visual perception, explains many reported perceptual biases, including the repulsive biases away from prior peak, which are counter-intuitive according to the traditional Bayesian view. This framework also offers a principle way to address the common criticisms of Bayesian models in perception, which argue that Bayesian models are lack of constraints. In the second study, I demonstrate that the idea of efficiency, coupled with a few assumptions, allows us to make quantitative predictions on the functional architecture of the grid cell system in rodents. One such prediction is

that the spatial scales of grid modules should follow a geometric progression, importantly, with the scaling factor to be close to the square root of transcendental number e ~ 1.6. Such zero-parameter predictions closely match the data reported in recent neurophysiological experiments. The theory also makes several other predictions, some of which have been confirmed by the data. This study suggests that achieving efficiency computation may also apply to neural circuits involving a high-level cognition, i.e. representation of space. In the third study, I analytically derive a generic connection between mutual information and Fisher information. This clarifies an important theoretical issue which has been misunderstood in previous neural coding literature. Additionally, it provides some powerful signatures of the Efficient coding hypothesis, which could guide future experimental tests. Together, the results presented in this thesis suggest that achieving efficient computation serves as a basic design principle which generalizes across neural systems processing low-level and high-level cognitive functions.

Contents

List of Figures

Chapter 1

Introduction

The human brain is a complex system which has attracted much endeavor to understand how it works. The scientific investigations have traditionally been dominated by the experimental approach, e.g. experimental psychology and neuroscience. In such experiments, one typically manipulates a single variable at a time. By observing how the target variable changes, one could potentially obtain some understanding about the underlying process. Although a great deal of the knowledge about the brain has been accumulated by this approach, one limitation is that it falls short when dealing with the huge dimensionality of the underlying parameter space, and when facing the complexity of computational machinery in the brain.

Starting from mid-20th century, scientists have gradually realized that, to understand the computations which gives rise to the various kinds of functions of the brain, one would need to rely on the language of mathematics and physics. The

use of mathematics and physics has many advantages and I shall only list a few of them. First, such language makes it possible to precisely describing the process of information processing in the neural systems. Second, one may gain some further insights of such process by analytical investigations based on the formulated mathematical model. Such insights, including these which may first appear to be counter-intuitive, are otherwise difficult to obtain by, *e.g.* reasoning via human language. Third, it could generate quantitative predictions to guide further experimental investigation, and increase the chance of observing interesting phenomena. The use of mathematics in neuroscience has already enjoyed some remarkable success back to several decades ago, such as the landmark results of Hodgkin-Huxley equations[93].

The approach of studying the brain by exploiting knowledge from mathematics and physics have now lead to the field of *computational neuroscience* [48]. Now there seems to be an increasing agreement in term of that coupling the computational approach with experimental approach offers the best promise to understand the brain. Most existing work in the field of computational neuroscience deals with computational models, which partly follows the tradition and style of the seminal work of Hodgkin-Huxley equations [93]. Arguably, only a small portion of the field deals with theories of the neural processing. Although the distinction between theory and model is often considered to be fuzzy, I believe that the difference is real. As articulated nicely by Charles Stevens, one key distinction between the theory

and model is that "*(...) models describe a particular phenomenon or process, and theories deal with a larger range of issues and identify general organizing principles*" [166]. While computational models are useful in providing quantitative descriptions of the underlying biological process, they only shed limited general insight into the designing principles of the brain. Overall, the computational modeling approach often addressed the "*how*" question. However, a more comprehensive understanding of the brain also requires asking the "*why*" question, which is typically the main focus of theories in neuroscience.

Theories in neuroscience, although so far less well developed, have nonetheless helped advanced our understanding of the brain in many ways. Notably, there have been two lines of theoretical investigations which have substantial impact on understanding of functions the brain, particularly perception. The first line of theories promotes the idea of treating perception and cognitive functions in general as an inference process, more specifically Bayesian inference(e.g. [103]). This route could be traced back at least to Helmholtz (1866) [90]. In another line, neural computation is formulated as a process of representing, recoding, and transmitting information. Specifically, the Efficient Coding hypothesis by Attneave (1954) and Barlow(1961) [10, 4] has sparkled a lot of following up investigations on early sensory processing.

The research presented in the thesis is mainly driven by the theoretical approach. The power of a good theory lies in its ability to explain various sets of data which

otherwise appear to be unrelated, and the ability to predict what would be observed in different situations. To this end, I have tried to test the developed theories using a wide range of data, ranging from psychophysical measurements of human behaviors to the neurophysiological measurements of neural populations in rodents.

1.1 Main hypothesis

The main hypothesis pursued in this thesis is that *the brain performs computations efficiently.*

Various considerations jointly point to the concept that the computation in the brain should be efficient (e.g. [113, 35, 112, 16, 116]). First of all, the computations performed in the brain is costly in terms of energy. One remarkable observation is that the brain, which typically consists 2% of the mass of the body, consumes 20% of the resting metabolic energy [35]. Every spike and every synaptic event in the brain requires energy [112, 116]. Second, consider the large variety of tasks which the human brain can perform. These range from perception, navigation, identifying objects, finding partners to solving math problems and thinking, writing. This is particularly remarkable given the size and the limited amount of possible energy consumption of the brain[112, 116]. To complete these various tasks successfully, the brain has to somehow use the available information and energy efficiently. Third, the survival pressure in evolution would tend to push the computations in the brain to a more efficient regime in order to increase fitness. Putting together, the survival

pressure and limited resource (e.g. energy constraint, information constraint) faced by the brain may have pushed the neural system toward the regime of being efficient during evolution.

Historically, the concept of *efficiency* has played an important rule in formulating theories of the brain computation. To appreciate this, let us briefly examine these two influential lines of theories mentioned above. In the case of treating perception as Bayesian inference, it could be seen as a result of coupling *efficiency* with statistical estimation theory. While there are many possible ways of doing statistical inference, the Bayesian inference representing the most efficient one. When *efficiency* is coupled with Information theory [153], the Efficient Coding hypothesis emerges [10, 4]. The appearance of the same concept in these two major branch of theories should not be considered as surprising, given that *efficiency* provides a fundamental, yet biologically well-grounded ingredient when formulating theories of the brain computation.

In this thesis, building upon previous research, I test this main hypothesis in both low-level and high-level cognitive functions of the brain, focusing on visual perception and spatial navigation. I shall demonstrate that the idea that the brain performs computation *efficiently* can explain a wide range of experimental observations, both in terms behavioral and neural measurements. Furthermore, it may provides a promising way to bridge different levels of observations.

1.2 Structure of the thesis

In Chapter 1, I elaborate the concept of *efficiency* and why it may be relevant for understanding the computations in the brain.

Chapter 2 aims to prepare the readers a minimal background for the materials presented in the three following chapters, without thorough reviews of these topics. In more details, it introduces i) two prominent hypothesis for understanding perception, namely *Efficient coding* and *Bayesian inference*; ii) the basics of the neurophysiological underpinning of spatial representation in rodent's brain; iii) a quick primer on Information-theoretic quantities, including the basic properties of mutual information and Fisher information.

Chapter 3 develops a general framework for understanding perception. While previous works on perception have either focus on the encoding or the decoding aspects, the presented framework integrates the idea of *Efficient coding* and *Bayesian decoding* into a model of perceptual behaviors. I shall demonstrate that this framework naturally accounts for various puzzling psychophysical observations reported previously. The examples involved are from visual perception. This Chapter is large part identical to a manuscript which has been submitted for consideration of publication.

Chapter 4 applies the idea of Efficient coding to a high cognitive function, i.e. neural coding of animal's self-location during spatial navigation. Although previous investigations demonstrate that Efficient coding may serve as a fundamental

principle for understanding early sensory processing, however, it is largely unknown whether such principle would also be relevant when studying high level cognitive functions. In this chapter, I demonstrate that the notion of "efficiency" quantitatively predicts the neurophysiologically observed functional architecture of the rodents' grid cells, which form an representation of the space. This Chapter is large part identical to a manuscript which has been submitted for consideration of publication.

In Chapter 5, I develop some analytically tools which clarifies the relationship between mutual information and Fisher information - two widely used quantities in neural coding. This important relationship has been misunderstood in previous works. Furthermore, the results provide some powerful tests of Efficient coding for future experiments. This Chapter is large part identical to a manuscript which has been submitted for consideration of publication.

The final Chapter summarizes the contribution of the thesis, discusses open questions and future directions which are likely to be fruitful.

Related publications and presentations

The work in Chapter 3 was conducted jointly with Alan Stocker. Part of this work was published in Neural Information Processing System (NIPS, 2012) meeting as a conference proceeding [184]. Part of this work was also presented in Computational and System Neuroscience meeting (CoSyNe, 2013) and Annual Meeting of Vision Sciences (VSS, 2014) as a poster. This work was also presented orally in

Models in Vision (Modvis, 2014) and Optical Society Vision meeting (2014). This work won *the best student poster award* in VSS, 2014. This manuscript which has been submitted for consideration of publication.

The work in Chapter 4 was conducted jointly with Jason Prentice and Vijay Balasubramanian. Part of this work was presented in Computational and System Neuroscience meeting (CoSyNe, 2013). A version of this work has appeared as a format of preprint on arxiv [182]. This manuscript which has been submitted for consideration of publication.

The work in Chapter 5 was conducted jointly with Alan Stocker. This manuscript has been submitted for consideration of publication.

Publications and presentations not included in this thesis

X.-X. Wei & A. A. Stocker. Bayesian inference with efficient neural population codes. In Artificial Neural Networks and Machine Learning –ICANN 2012, pages 523–530. Springer, 2012.

J. Jacobs, C.T. Weidemann, J. F. Miller, A. Solway, J. F. Burke, X.-X. Wei, N. Suthana et al. Direct recordings of grid-like neuronal activity in human spatial navigation. Nature Neuroscience, 16(9):1188 –1190, 2013.

X.-X. Wei, P. A. Ortega, and A. A. Stocker. Perceptual adaptation: Getting ready for the future. Annual Meeting of Vision Sciences (VSS), 2015. Abstract. Winner of the *student travel award*. Winner of *the best student poster award*.

Chapter 2

Background

Due to interdisciplinary nature of the research presented in this thesis, a brief introduction for each topic seems to be desirable. In broad stroke, the research are related to four different areas – Efficient coding, Bayesian inference, spatial cognition, Information theory. Below I shall present the basics of each topic, with the goal to facilitate the readers' understanding of the following Chapters. Readers who are already familiar with these topics should feel free to skip some of these sections.

2.1 Efficient coding

One important hypothesis which drives the research on perception, in particular early sensory procession, is the Efficient Coding Hypothesis. Efficient coding was pioneered by Attneave (1954) and Barlow(1961) [10, 4]. These proposals were partly

inspired by the seminal work of Shannon on information theory [153]. One key contribution Atteave and Barlow brought into the field of neuroscience is the idea that the mathematical framework of information theory may be relevant for understanding the brain, and it may shed light on the strategy by which the neural system process information.

Specifically, Barlow's original version of Efficient coding concerns about how information should be encoded in the sensory system, particularly in early visual system. Barlow's idea is that the sensory signal should be recoded in the *most economical* way [10]. One particular way to increase efficiency, which is what Barlow emphasized, is to make the response of the two outputs to be more independent, i.e. *redundancy reduction*. The idea is simple. Redundancy between the output units would lead to a decrease in terms of the information transmitted. Thus, if the neural processing could somehow eliminate such redundancy, the efficiency of the representation should be improved.

Many versions of Efficient coding theories were developed afterwards which follow this idea of the *efficient* information transmission in the neural system. One notable example is the proposal by Linker that the neural system should be organized in a way such that the mutual information between the input and output should be maximized, i.e. *InfoMax* [121]. Importantly, Linsker also investigated how biologically plausible learning rules might give rise to such efficient representation [122, 123].

One may wonder why there are many seemingly different versions of Efficient coding theory co-existing in the literature. In my opinion, one reason is that different Efficient coding theories typically make different assumptions on the noise structure of the system [154]. For example, in Barlow's original proposal, consideration of the neural noise is not included [10]. In this situation, redundancy reduction is a desirable goal. Linkser's *InfoMax* could be related to the *redundancy reduction* proposed by Barlow in the zero-noise limit. In some models, Gaussian noise are assumed (e.g., [3, 179]). In the presence of certain type of noise, certain amount of redundancy could actually be advantageous [3, 179, 55].

Another reason is that different theories may impose different constraints on the overall resources and different objective functions. For instance, transmitting the same amount information using the fewest number of spikes is, in general, a different goal comparing to using the fewest number of neurons. These two objectives are also different from the objective of using the smallest number of simultaneously active neurons. These different assumptions would lead to different optimal configurations of neural code. Interestingly, later work on sparse coding was inspired by consideration of representing the input using a small number of active neurons [134, 135]. Some have argued that the optimal neural code should explicitly quantify and take into account of the amount of energy consumption [119, 112, 7, 6, 116]. Some others have argued that one should use the reconstruction error rather than the amount of information transmitted as the criteria(e.g., [177, 181, 149]).

Despite these different formulations, there seems to be one general agreement among these theories, which is that the neural system should exploit the statistical structure of the surrounding environment [63, 155]. This arises because, fundamentally, the amount of information transferred by a noisy channel, as well as the reconstruction error, crucially depends on the probability distribution of the input [153]. Because the input to the sensory system will inevitably shaped (at least partially) by the statistics of the environment, a good design of the sensory system would have to take such statistical regularities into account.

Substantial efforts have been devoted to test the Efficient coding hypothesis experimentally. Such investigations have lead to much success in early visual (e.g. [111, 46, 6, 134, 59, 8, 100, 76]) and auditory processing (e.g., [120, 157]). However, challenges remain. First, most studies have been done on early sensory system, and it is presently unknown whether these results would generalize to relatively higher cognitive area or not. Second, from a theoretical point of view, many previous research have tried to predict the tuning property of individual neuron (e.g., [111, 127, 181]). However, arguably more desirable tests, which would shed more light on the neural processing, should be tests at the neural population level. Third, most previous work have focused on the predictions in terms of neurophysiological aspects, while the connections to behavior are far less common (but see [9, 76, 92]).

2.2 Bayesian inference

Besides Efficient Coding, another major hypothesis which has profound impact on the understanding of the perceptual process is the proposal of treating perception as Bayesian inference (e.g. [103]). The basic idea is that perception is not simply taking pictures of the environment like the way a camera does, rather it involves active interpretation of the raw sensory inputs. Thus, perception could be better viewed as an inference process. Holmherz already realized this concept in the 19th century [90].

To appreciate the perspective of treating perception as an inference process [103], it is useful to realize that there is virtually always ambiguity in the sensory input. Consider the example of vision, in which 3-d environment is mapped to 2-d retinal image. Some information becomes lost inevitably in such mapping. Also consider the fact the noise is not only present in the stimulus itself, but it is also ubiquitous along the sensory processing in the brain [60]. The presence of noise leads to ambiguity when interpreting the sensory observation, and in principle could result in many different interpretations of the same stimulus.

The noise in the perceptual process imposes a fundamental challenge for perception (not only for visual perception), namely, how does the perceptual system select the specific way to interpret the input? An appealing hypothesis is that perception involves *efficient* interpretation based on limited information gathered by the sensors. As it turns out, Bayesian inference provides such an efficient way to

interpret the sensory observation.

Bayes' rule

To understand how Bayesian inference works, consider the case of judging the speed θ of an moving object. The perceptual system takes some observations of speed of the moving object, which will be termed as "*data*". Note that the process which generates the *data* is typically noisy, in the sense that the mapping from a particular speed θ to the *data* is not one-to-one, rather such dependence can be summarized as a conditional probability distribution $P(data|\theta)$. The problem is that, once the *data* is known, how to infer the underlying speed of the object? Bayesian inference tells us that one should use both the prior belief on the speed of the moving object and the evidence gathered from the observations to perform the inference. Furthermore, crucially one should combine these two source of information using Bayes rule (Bayes &Price, 1763). Mathematically, Bayes' rule could be written as

$$P(\theta|data) = \frac{P(\theta)P(data|\theta)}{P(data)} \tag{2.1}$$

This rule is the core of Bayesian inference. The term on the left-hand side $P(\theta|data)$ is the posterior distribution on θ given the *data*. On the right-hand side, $P(\theta)$ is typically referred as prior distribution on θ, which summarize the prior belief on θ before the observation. The term $P(data|\theta)$ is called the likelihood function

of θ, which summarizes the evidence gathered from the *data* . It is important to emphasize that although $P(data|\theta)$ itself is a probability given θ, but with respect to variable θ while fixing *data*, it is a function of θ, because its value varies with θ. Therefore, it is appropriate to call it a likelihood function on θ, rather than a "likelihood distribution". Finally, the term $P(data)$ is a normalization factor which guarantees that $P(\theta|data)$ is a proper probability distribution.

A common goal for Bayesian models is to derive the posterior distribution on θ, i.e. $P(\theta|data)$. This is, again, proportional to the product of the prior distribution and the likelihood function. To get the posterior distribution, naturally one has to obtain the prior belief and the likelihood function first.

To calculate the likelihood function precisely, in principle the full knowledge of how the *data* are generated from θ is required. Such information critically depends on the structure of the model (sometimes called "generative model" [48]). Of course, the complexity of the problem varies depending on the structure of the model. Chapter 3 has a detailed discussion on this issue, where I shall propose a principled way to specify the likelihood function for certain perceptual inference problems.

How to specify the prior distribution? In practice, the selection of prior distributions is often done based on computation convenience. For example, a prior distribution has often been chosen to be flat or Gaussian. Alternatively, a seemingly more reasonable way is to pick up a "reference prior" [14] based on some information criteria. The question of the selection of prior distribution is an active

research topic in statistics [14, 101]. For perceptual inference, there may be principled ways to specify the prior. Consider the speed example we mentioned again. In this case, it seems reasonable to assume that such prior belief should depend on our sensory experience in the past. This means that the prior distribution should reflect the statistics of given variables in natural environments. Again consider the example of the perceived speed. The prior may be thought as the statistics of the speed in natural environments, which determines our long term sensory experience on speed [169]. However, this is just a first-order approximation. In general, the prior may depends on various other factors, e.g. context, individual, and so on.

Loss function

In many cases, computing the posterior distribution is not the end of the task, because a particular Bayesian estimator may be required. For example, in the context of speed estimation, a particular estimation of the speed of the object has to be obtained, while the posterior distribution itself is not enough to specify the percept. Technically, the mapping from the posterior to the Bayesian estimator could be done by first assuming a loss function, and then constructing the corresponding optimal estimator according to such loss [40]. Common choice of the loss functions involves the family of L_p loss functions with p can be chosen to be different values. For squared error loss ($p = 2$), the resulting Bayesian estimator is the mean value of the posterior distribution. For absolute error loss ($p = 1$), the corresponding

estimator is the median of the posterior distribution. The MAP estimator, i.e. the mode of the posterior distribution, is obtained under $0-1$ loss ($p=0$). Not too much is known on the specific loss functions used by the perceptual system. We will return to this point in Chapter 3.

2.3 Neural representation of physical space

As we navigate around, we often have a sense of where we are in space. Actually, knowing the self-location relative to the environment is a fundamental ability for the purpose of survival. How does the brain support such ability? Tolman (1948) proposed that the brain should be able to maintain a "cognitive map" of the physical space [174]. The discovery of place cells in rodent hippocampus around 1970s has triggered a lot of following up research on the neural basis of spatial map in the brain [132, 133]. The most salient property of place cells is that its firing activity is sparse in space. Individual place cell typically fires when only animal is within a particular spot in space [132], although some place cells fire in multiple spots in space. As we know more about the place cells, it becomes clear that the firing activities of the place cells are also correlated with many other factors besides the animal's spatial location, including odor, recent experience, time and others [138, 2, 136]. The interpretations of these multi-perplexing responses are still subjected to debates.

Although the response of place cell are strongly correlated with space, it is un-

clear whether the spatial information is originated within hippocampus or inherent from other brain areas. The discovery of grid cells in dorsal-medial Entorhinal Cortex offers some new insight to this question[86, 75]. Now it appears that place cell may, at least partially, inherent the spatial information from grid cells. Place cells, grid cells, together with heading direction cells[172] and boarder cells[159] may consist the neural underpinning of our sense of space during navigation. Studying the properties and functions of these cell may provide us a unique chance to uncover how the space is represented in the mammalian brain, and how spatial maps in the brain support navigation behaviors, which are critical for the survival of mammals.

Grid cells

In 2004, Fyhn et al., discovered that in dorsal band of Entorhinal Cortex (EC) of the rat's brain, neurons typically show tuning preference for space, similar as the place cells [75]. However, unlike place cells, these cells typically fire in the multiple spatial locations. In 2005, a following up paper by Hafting et al., demonstrated that, surprisingly, the spatial firing fields of the cells in dmEC lies on a triangular grid [86]. These cells are thus termed as "grid cells". Because the highly regular firing pattern of the grid cells, it is immediately suggested that the grid cells may encode a metric of the space [86, 129]. Grid cells are also found in pre-and parasubiculum [19], two brain regions next to EC. Although the response pattern of grid cells can be partially manipulated by cues in the environment [86], overall the major factor

which determines the response of the grid cells response appear to be the animal's location in space.

Why are grid cells not discovered in the studies before Hafting et al. (2005) [86]? One major reason seems to be that previous experiments have used smaller testing rooms, which were not enough to reveal the lattice structure of the grid. In a small environment, typically only one or even zero firing field of individual grid cells could be observed. However, in larger recording rooms, the pattern of the grid become visually apparent. The grid is particularly evident when the two dimensional autocorrelation map of the grid firing map is calculated which effectively reduce the noise in the original firing rate map by averaging [86]. Originally found in rats, the grid cells are later discovered in mice[74], and in bats [187]. There are some indirect evidence from fMRI signal suggesting that grid cells may also exist in humans [54]. Recently, my colleagues and I have reported the first direct observation of grid-like response in human brain by analyzing data from single neuron recording of human epilepsy patients, while they were performing a virtual navigation task [94].

The response pattern of individual grid cell can be characterized by three parameters, the spacing, the orientation, and the spatial phase. Locally, the grid cells in rodents share similar spacing and orientation [86, 164]. The spatial phase seems to be shifted randomly such that nearby cells do not have nearby phases[86]. Therefore, there seems to be no topographical relationship in the spatial phase. Interestingly, the grids in EC have different scales manifested in the spacing of

the grids. Furthermore, the scales increase systematically along the dorsal-ventral axis [86, 26]. At the dorsal most, the spacing of the grid is about 50cm, while at the 75% of the dorsal-ventral axis, the spacing of the grid can be several meters to 10 meters, according to the recording when rats running on a 18 meters linear track[26].

One particular important property of grid cells is that they are organized in discrete structure [164]. By a fine sampling of grid cells along about half of the dorsal-ventral axis of EC, Stensola et al. (2012) demonstrates that grid cells could be clustered based on the scale, orientation and ellipticity [164]. The cells within one cluster share the same scale, orientation and ellipticity. The individual cluster is termed as a module. These authors found up to 5 modules within individual animal. However, because the recording was only done up to 50% of the dorsal ventral axis of EC, more modules should be expected in whole EC. A simple linear extrapolation suggests that the number of the grid cell modules in rodent EC should be ~ 10. Strikingly, the data also suggest that the grid scales follow a geometric progression. In this particular data set, the scaling factor was found to be ~ 1.42, while in a previous study, the scaling factor was reported be ~ 1.7 with a relatively smaller sample size [11].

The grid cells have attracted many computational investigations since it is discovered. Most research have focus on the mechanisms and algorithms of how the grid-like response could be generated. Existing models have exploiting mechanisms

of pattern formation[176] in attractor network [72, 28, 20] and oscillatory interference [30, 89], as well as spike rate adaptation [106].

While these computational models of grid cells aim to address the *how* question, the question of *why* the grid cells firing pattern should be as observed remain mysterious. In Chapter 4, I ask the fundamental issue in terms of why it is desirable to use the grid code to form a representation of the space. I show that the idea of *efficient* processing of spatial information quantitatively accounts for the functional architecture of the grid cells observed in the rodent's brain.

2.4 Mutual information and Fisher information

In this section, I introduce two important Information-theoretic quantities which are used frequently in this thesis, namely mutual information and Fisher information. Note that mutual information comes from Information theory, while Fisher information has a fundamental root in statistics.

2.4.1 Mutual information

Information theory concerns the communication of information [153]. A basic quantity in Information theory is *entropy*. *Entropy* characterizes the amount of uncertainty associated with a random variable. For a discrete random variable X with distribution $p(x)$, its *entropy* can be defined as

$$H[X] = \sum p(x) \ln p(x)$$

.

For a continuous random variable X with density $p(x)$, its (differential) *entropy* can be defined as

$$H[X] = \int_X p(x) \ln p(x) dx$$

.

Mutual information quantifies how much information of one random variable is contained in another random variable. Formally, mutual information could be expressed as

$$I[X, Y] = H[X] - H[X|Y].$$

Note that following this definition, we can also consider

$$I[Y, X] = H[Y] - H[Y|X].$$

It can be verified that the following is true

$$I[X, Y] = I[Y, X].$$

For two continuous random variables X, Y, with joint probability density $p(x, y)$, mutual information can be computed explicitly as

$$I[X,Y] = \int_Y \int_X p(x,y) \ln \left(\frac{p(x,y)}{p(x)p(y)} \right) dxdy.$$

Two important facts. First, $I[X,Y]$ is always non-negative. Second, $I[X,Y] = 0$ is equivalent to the independence of X and Y.

Information theory has profound influence in many scientific fields, including neuroscience. In many neuroscience problems, one could treat individual neurons or neural populations as noisy communication channel(s) and study the information transmitted in such channel(s). For instance, denoting the stimulus variable as s, and the neural response as r, one could compute the mutual information between the stimulus and the response as

$$I[r,s] = H[r] - H[r|s].$$

This means that the mutual information is, technically, the difference between the entropy of the response and the entropy of the response given a particular stimulus. Conceptually, this captures the ratio between the volume of the response space and the average volume of the response given a stimulus.

Alternatively, one could compute the mutual information as

$$I[s,r] = H[s] - H[s|r].$$

The interpretation for this expression is that, mutual information could also be

computed as the difference between the entropy of the stimulus and the entropy of the stimulus given a response. Conceptually, this is related to the ratio between the volume of the stimulus and the average volume of the stimulus given a response. Thus it quantifies how well a response r could tell about the stimulus s.

Recall the mathematical fact that $I[s, r]$ is *always* identical to $I[r, s]$, although it should be apparent from the above discussions that the conceptual interpretations of these two quantities could be quite different. Practically, the mathematical equivalence of these different expressions offers two choices for computing the mutual information. Depending on the problems, one expression is usually easier to work with compared to the other one. Unfortunately, sometimes both expressions are difficult to compute exactly, in which cases one would need to rely on further assumptions to work with these quantities.

2.4.2 Fisher information

Fisher information [67] is one central quantity in statistics, particularly in estimation theory [115]. Fisher information is sometimes referred as "information" in statistics [115]. Note that this should not be confused with mutual information defined above. Fisher information characterizes the amount of information that an observation (or measurement) m carries about an unknown parameter θ given the statistical relationship between m and θ. If both θ and m are scalar, Fisher

information can be defined as

$$J(\theta) = \int \left(\frac{\partial \ln p(m|\theta)}{\partial \theta} \right)^2 p(m|\theta) dm. \tag{2.2}$$

In this expression, $p(m|\theta)$ should be treated as the likelihood function on θ. $\ln p(m|\theta)$ represents the log-likelihood function, which is often called "support curve" [58]. The slope of the support curve is the "*score*", which represents how sensitively the likelihood function depends on the parameter θ. It is important to emphasize that while the likelihood function depends on a particular observation m, Fisher information does *not*. The reason is that the dependence on m is integrated out by definition. Thus, it is appropriate to interpret Fisher information as a measure of the expected sensitivity with respect to each value of the parameter θ, which is fully determined by the encoding model which specifies the relationship between the random variables m and θ. In some sense, Fisher information defines a metric in the space of θ.

As a remark, there is another way to define Fisher information

$$J(\theta) = -\int \frac{\partial^2 \ln p(m|\theta)}{\partial \theta^2} p(m|\theta) dm. \tag{2.3}$$

It is straightforward to check that these two definitions are equivalent. These definitions only apply when θ is a scalar. If θ is a vector, one can define a corresponding Fisher Information matrix [115, 1].

Fisher information has many interesting properties. For the purpose of this thesis, I shall only introduce a few of them.

Cramer-Rao bound

Perhaps the most well-known result related to Fisher information is the Cramer-Rao bound [41, 139]. Cramer-Rao bound states that, under certain regularity conditions, Fisher information sets an lower bound on the variance of any *unbiased* estimator $\hat{\theta}$. Formally, it could be expressed as

$$Var(\hat{\theta}) \geq \frac{1}{J(\theta)}. \tag{2.4}$$

In general, Cramer-Rao bound could not be reached. It can only be tight for special kinds of statistical models. I shall come back to this point later in Chapter 5, where the conditions to make Cramer-Rao bound tight are discussed in some more details. Intuitively, Cramer-Rao bound means that the quality of encoding, quantified by the Fisher information, set a physical limit on how precise any *unbiased* estimator can be.

For *biased* estimator, the corresponding Cramer-Rao bound turns out to be

$$Var(\hat{\theta}) \geq \frac{[1 + b'(\theta)]^2}{J(\theta)}, \tag{2.5}$$

where $b(\theta)$ represents the bias of the estimator $\hat{\theta}$. It then follows that the MSE

(mean square error) of estimator $\hat{\theta}$ must satisfy

$$MSE(\hat{\theta}) \geq \frac{[1 + b'(\theta)]^2}{J(\theta)} + b(\theta)^2. \tag{2.6}$$

It is useful to point out that the MSE of a biased estimator could be smaller than $\frac{1}{J(\theta)}$ which defines a lower bound for any *unbiased* estimator. Although many might have the intuition that an unbiased estimator is advantageous compared to a biased estimator, this result suggest that, counter to that intuition, having a bias in the estimation could be actually desirable in certain situations.

Invariance of Fisher information

The square root of Fisher information $J(\theta)$ has a property of invariance, i.e.

$$\sqrt{J(\theta)}d\theta = \sqrt{J(\tilde{\theta})}d\tilde{\theta}, \tag{2.7}$$

where $\tilde{\theta}$ is a re-parameterization of θ. As a corollary, the integral $S = \int \sqrt{J(\theta)}d\theta$ is invariant with respect to any re-parameterization of θ. Under these notations, $f_J(\theta) = \frac{\sqrt{J(\theta)}}{S}$ behaves like a probability density. $f_J(\theta)$ is known famously as Jeffreys prior [97], which is a widely used non-informative prior in Bayesian statistics [101]. As a remark, the integral $\int J(\theta)^p d\theta$ when taking p other than $\frac{1}{2}$ is not invariant with respect to the re-parameterization of θ. In this sense, $\sqrt{J(\theta)}$ is special.

Relationship to psychophysical and neural measurements

Fisher information has widely applications in many scientific fields. In the extreme, It has even been argued that Fisher information can provide a unification of many area of science[71]. In this thesis, I shall focus on its possible applications in terms of understanding the information processing in the brain. Let me start by noting that Fisher information has a nice relationship with respect to the most commonly taken psychophysical measurement, i.e. discrimination threshold. It has been well-established that [152, 151] Fisher information sets an lower bound on the discrimination threshold $(\dot{\theta})$ in fine discrimination tasks:

$$d(\theta) \geq C_\alpha \frac{1}{\sqrt{J(\theta)}},$$

where C_α is a constant determined by the specifics of the psychophysical procedure.

Fisher information can also be used to assess how much information a certain neuron (or neurons) carries about a particular stimulus dimension. Consider a Poisson neuron with a smooth tuning curve $f(\theta)$. In this case, the Fisher information has a nice close-form expression

$$J(\theta) = T \frac{f(\theta)'^2}{f(\theta)},$$

where T represents the length of the integration time. There are several basic insights from this expression. First, both the firing rate and the slope of the tuning

curve of a Poisson neuron are important in terms of the Fisher information the neuron's response carries. Second, the neuron carries most Fisher information at the flank of its tuning curve rather than the peak. Third, the Fisher information scales linearly with the integration time and the gain of the neuron.

Fisher information and mutual information have intriguing relationships. On one hand, by definition, Mutual Information and Fisher information are quite different. Conceptually, Fisher information quantifies the *local* information, while mutual information is a *global* measure. On the other hand, these two measures are also closely related, as we will discuss in details in Chapter 5.

Chapter 3

Bayesian observer model constrained by Efficient coding explains "anti-Bayesian percept"

3.1 Introduction

Perception involves two important stages of processing: 1) the representation of incoming sensory information, and 2) the interpretation of that representation to form a percept. Two prominent hypotheses have separately guided our understanding of these two processing stages, but each has limitations when considered alone. The *Efficient Coding Hypothesis* argues that neural resource limitations lead to efficient sensory representations that are optimized with regard to the specific stimulus

statistics of the natural environment [4, 10]. This hypothesis can explain several key features of neural coding in early sensory areas (*e.g.* [134, 46, 120]), but it does not specify how these coding characteristics can give rise to important aspects of perceptual behavior such as perceptual biases. In contrast, the *Bayesian Hypothesis* posits that perception is an act of unconscious inference that interprets the noisy sensory representation in the context of prior knowledge about the world [90, 44, 103]. This hypothesis provides a normative explanation for many aspects of perceptual and sensorimotor behavior (*e.g.*, [104, 169, 178, 98]), but it has been criticized for using arbitrary model specifications in order to explain psychophysical data [99, 21]. Here we unify ideas of Efficient coding and Bayesian inference into a new model of perceptual behavior. Specifically, we propose an Bayesian observer model that is constrained by assuming an efficient representation of the sensory input.

Two key components define a Bayesian observer: the prior belief that reflects the observer's expectation about how frequently a certain stimulus value occurs, and the likelihood function that captures the encoding accuracy in the sensory representation of the observer. Previous studies have proposed independent constraints on either the prior belief based on natural (*e.g.*, [175, 83]) or learned (*e.g.*, [96, 104]) stimulus statistics, or the likelihood function based on natural stimulus uncertainties(*e.g.*, [78, 29]) or neural physiological tuning characteristics (*e.g.*, [169]), but not both. In contrast, our new model formulation jointly constrains both the prior belief and the likelihood function by assuming that the sensory representation as

well as the interpretation of the sensory evidence is optimized with regard to the stimulus statistics of its sensory environment. Thus, we can specify a Bayesian observer model for any stimulus variable with known natural statistics.

We validated our framework by formulating observer models for two perceptual variables for which the natural statistics are known, visual orientation and spatial frequency. The models make a number of distinct and rather surprising predictions; *e.g.*, that percepts are frequently biased away from the peaks of the prior, a prediction that seems at odds with the standard Bayesian view. We demonstrate that the predictions are well matched by data from several studies reporting measured biases in perceived visual orientation and spatial frequency under different levels and sources of uncertainty. That includes biases that are seemingly "anti-Bayesian" [23]. Our results demonstrate that by combining the ideas of Efficient coding and Bayesian decoding, we can formulate well constrained observer models that can account for perceptual behavior that has not been explained before. Some earlier version of this work has been previously presented [184].

3.2 Results

We model perception as a probabilistic encoding-decoding process (Fig. 3.1a) [169]: The presentation of a stimulus with a single value θ elicits a noisy sensory measurement m (encoding), based on which the observer then generates an estimate $\hat{\theta}(m)$ that represents the perceived stimulus value (decoding). We combine two general

assumptions in order to define our observer model. First, we assume that encoding is efficient, *i.e.*, the sensory representation is optimally adapted to the natural stimulus distribution. Second, we assume that decoding is Bayesian and is based on an accurate (generative) model of the sensory process, *i.e.*, the observer's prior belief matches the true stimulus distribution and the likelihood function faithfully reflects the encoding characteristics. As a result, both the observer's prior belief and likelihood function are jointly constrained by the stimulus distribution (Fig. 3.1b). Thus, with the additional assumption about the observer's loss function (that states how costly some perceptual errors are for the observer), we can make quantitative predictions for the percept of a stimulus variable for which the natural stimulus distribution is known. In the following we show how to formulate the model and derive these predictions, and how they compare to measured psychophysical data.

3.2.1 Efficient coding and the likelihood function

We adopted a definition of Efficient coding that assumes that sensory encoding maximizes the mutual information $I[\theta, m]$ between the sensory measurement m and the stimulus variable θ with regard to the intrinsic uncertainty (internal noise) in the sensory representation [121]. The definition establishes a link between the probability distribution of the stimulus $p(\theta)$ and Fisher information $J(\theta)$ using a

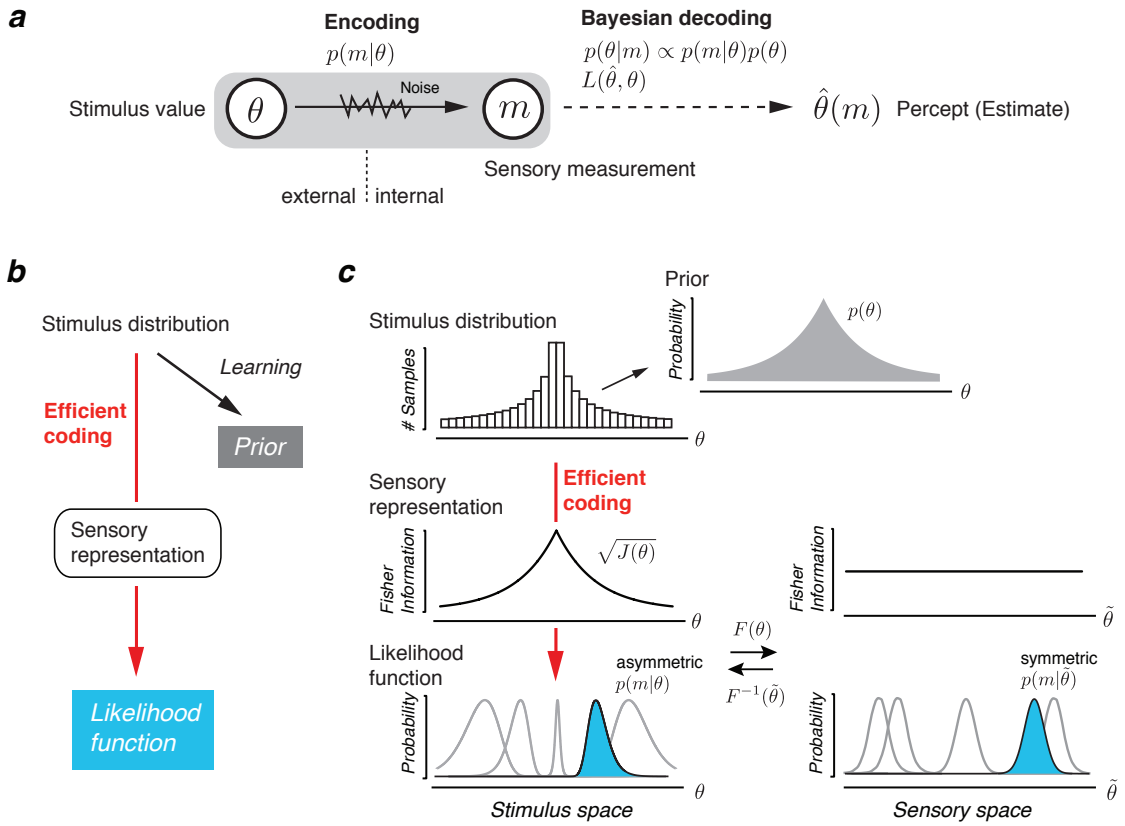

Figure 3.1: *Bayesian observer model constrained by Efficient coding.* a) We model perception as an encoding-decoding process. Encoding is characterized by the corresponding conditional probability distribution $p(m|\theta)$ of the sensory measurement m given a stimulus value θ. We assume encoding is governed by Efficient coding. We also assume that decoding is Bayesian based on an accurate generative model of the sensory process. The percept $\hat{\theta}(m)$ is then specified based on the posterior distribution $p(\theta|m)$ and a loss function $L(\hat{\theta}, \theta)$. b) Our assumptions imply that the Bayesian observer is constrained by the natural stimulus distributions: The prior belief is assumed to directly match the stimulus distribution (*e.g.* through learning), while the likelihood function is constrained by the stimulus distribution via Efficient coding. c) Example for an arbitrary stimulus distribution. An Efficient coding principle that maximizes mutual information implies that the encoding accuracy (measured as the square-root of the Fisher Information $J(\theta)$) matches the stimulus distribution. With some assumptions about the sensory noise characteristics the likelihood function is fully constrained by the Fisher Information. Likelihood functions for different sensory measurements m are shown to illustrate their heterogeneity across the stimulus space. Technically, the likelihood functions can be computed by assuming a symmetric noise structure (*i.e.*, symmetric likelihood functions) in a space in which the Fisher information is uniform (sensory space, characterized by the mapping $F(\theta)$), and then transforming those symmetric likelihood functions back to the stimulus space.

bound on mutual information [27, 127]. Assuming the bound is tight it follows that

$$p(\theta) \propto \sqrt{J(\theta)} \qquad \text{(see Methods for details).} \tag{3.1}$$

Fisher information $J(\theta)$ is a measure of encoding accuracy and reflects the amount of sensory resources that is dedicated to the representation of a certain stimulus value θ. Equation (3.1) provides an intuitive way of understanding Efficient coding: Sensory resources should be allocated according to the stimulus distribution $p(\theta)$ resulting in a more accurate representation of those stimulus values that occur more frequently.

Fisher information directly constrains the likelihood function, given our general assumption that the likelihood function faithfully reflects the encoding characteristics. But it is not sufficient to fully specify the shape of the likelihood function. An additional assumption about the noise structure is required. Let us consider a function $F(\theta)$ that maps the stimulus space to a new space in which Fisher information is uniform (see Fig. 3.1c). We refer to this space as the "sensory space" [1]. With our chosen Efficient coding constraint Eq. (3.1) the mapping $F(\theta)$ is defined as the cumulative of the stimulus distribution (prior) [111] (see Methods, Eq. (3.8)). Uniform Fisher information implies that the noise and thus the likelihood function is homogeneous. We introduce the additional assumption that the expected likelihood

[1] In reference to Gustav Fechner because discriminability, when measured in units of this space, is uniform [61].

function (*i.e.* averaged out over many trials) is symmetric around the stimulus value in the sensory space. A simple way to guarantee this is to assume, for example, the noise to be additive and symmetric (*e.g.* Gaussian as illustrated in Fig. 3.1c). For a given sensory measurement m the likelihood function in the stimulus space can then be obtained by simply applying the inverse mapping $F^{-1}(\tilde{\theta})$. As a result, the likelihood functions when formulated in stimulus space are typically asymmetric with a long tail away from the peak of the prior distribution.

Note that by formulating the Efficient coding in terms of Fisher information we were able to specify the likelihood function without having to assume specific details about the tuning characteristics of the underlying neural representation. We deliberately chose such formulation because it allowed us a more parsimonious yet also more general description of our Bayesian observer model. In fact, as we demonstrate later, neural populations with quite different tuning characteristics but equivalent distributions of Fisher information can represent equivalent efficient sensory representations that lead to similar Bayesian decoding characteristics.

3.2.2 General predictions of the framework

The tight link between the stimulus distribution, the encoding accuracy (*i.e.*, the Fisher information $J(\theta)$), and the shape of the likelihood function has important consequences for the resulting decoding characteristics of our Bayesian observer model. In particular, it makes two novel predictions with regard to perceptual bias

that are surprising and counter-intuitive from a standard Bayesian modeling point of view.

The first prediction concerns the effect of the likelihood asymmetry on perceptual bias. A Bayesian modeling approach that assumes a symmetric likelihood function predicts the percept to be biased towards the prior peak for relatively smooth prior distributions (Fig. 3.2a). The situation changes, however, if the likelihood function is asymmetric (Fig. 3.2b). Now, the asymmetry itself can lead to estimation biases (see also [170]). In our framework, the shape of the likelihood is asymmetric for any non-uniform stimulus distribution with a heavier tail pointing away from the prior peak. This shape typically results in a repulsive bias component that we refer to as the likelihood repulsion. Although the effect depends on the chosen loss function, it is remarkably robust for commonly used choices (see Fig. 3.7). The repulsive effect is further amplified when computing the expected bias over many measurements of a given stimulus value θ_0. The reason is that the distribution of these measurements in the stimulus space also follows the same asymmetry; *i.e.*, the noisy measurements and thus the position of the likelihood functions on each trial are, on average, also biased away from the true stimulus value θ_0. These observations suggest a nuanced account of perceptual biases as the net result of two bias components, one introduced by the likelihood asymmetry and one by the prior distribution. Because of the above link, we can precisely predict the net bias for known natural stimulus distributions. We find that under many conditions

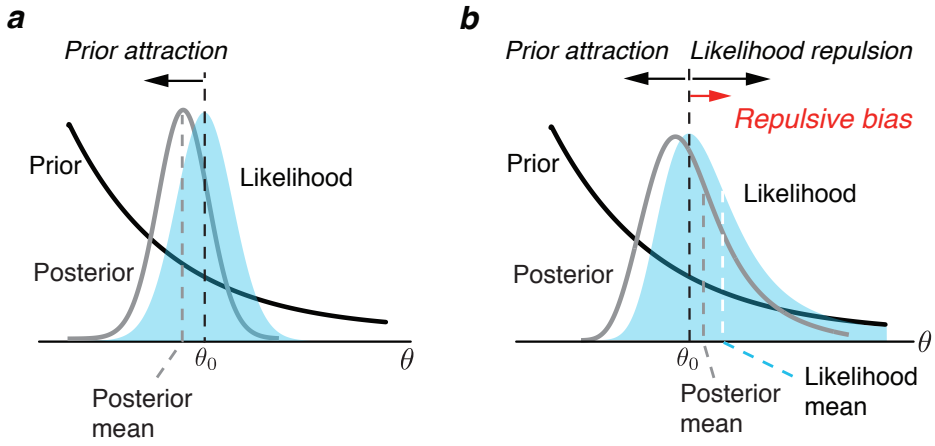

Figure 3.2: *Prediction 1: Bayesian perception can be biased away from the prior peak.* a) A standard Bayesian observer model that assumes a symmetric likelihood function typically predicts perceptual biases toward the peak of the prior. This "bias towards the prior" has been considered a fundamental characteristic of a Bayesian model. b) In our new Bayesian observer model, Efficient encoding promotes a non-homogeneous sensory representation that leads to an asymmetric shape of the likelihood function with a long tail pointing away from the prior peak. As a result, the estimate can be biased away from the prior peak. This is illustrated assuming the Bayesian estimate is determined by the posterior mean (squared-error loss function). Due to its asymmetry the mean of the likelihood function is away from the peak of the prior relative to the true stimulus value θ_0 (Likelihood repulsion). Although the prior still leads to an attractive shift of the posterior (Prior attraction), the net bias can be repulsive. Both examples are illustrated for the case of the median likelihood function (*i.e.* the measurement equals the stimulus value θ_0).

the model predicts that perception is biased away from the peak of the stimulus distribution (*i.e.* the prior belief). In particular, assuming small sensory noise only and a squared-error loss function (posterior mean) we can derive analytical solutions for the expected perceptual bias for arbitrary stimulus distributions (see Methods for details). The predicted bias is always repulsive if the prior distribution is well approximated by a monotonic function over the support of the likelihood function. This prediction is quite remarkable since the "bias towards the prior" has been considered a fundamental characteristics of Bayesian observer models.

The second prediction is that stimulus (external) and sensory (internal) noise differently affect perceptual bias. The difference emerges because our Efficient coding assumption generally imposes an inhomogeneous sensory representation that has a different metric than the physical space. Thus, although ultimately both sources of uncertainties are jointly reflected in the noise of the sensory measurement m their individual effects on the likelihood function are different because of the mapping function F (Fig. 3.3a). As a result, the same noise added at the stimulus level leads to a different likelihood function than the equivalent noise added at the sensory level, which results in a different bias.

Increasing sensory noise results in a likelihood function that is more asymmetric in the stimulus space because the additional uncertainty is mapped from the sensory space (where it is symmetric; *e.g.* Gaussian) to the stimulus space via the inverse mapping F^{-1}. Although the prior attraction increases due to the overall wider

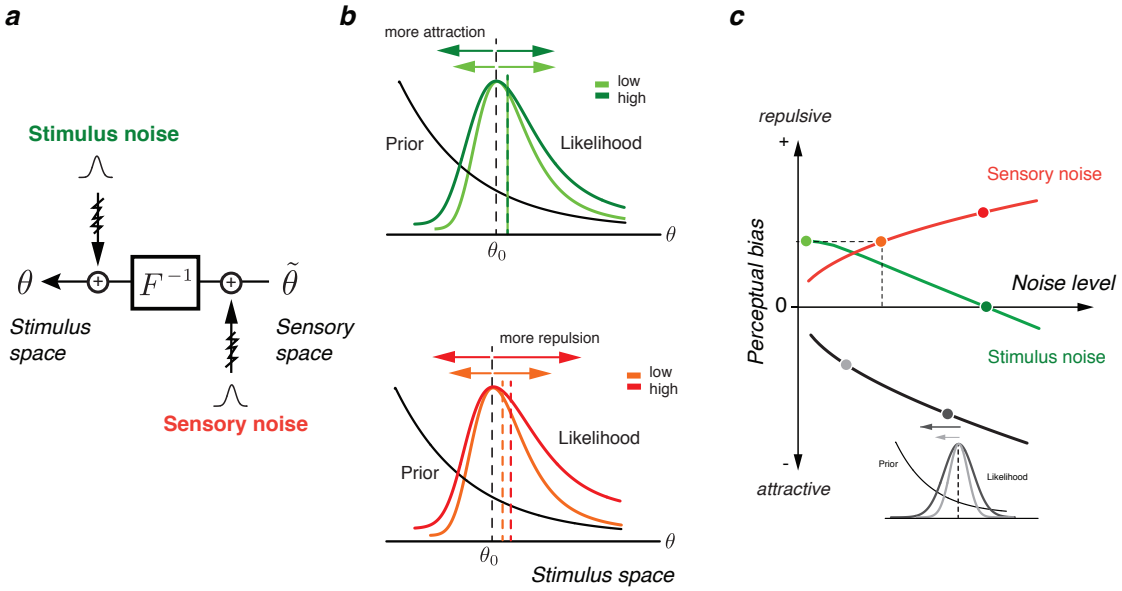

Figure 3.3: *Prediction 2: Stimulus (external) and sensory (internal) noise differentially affect perceptual bias.* a) Stimulus noise directly affects stimulus uncertainty and thus the likelihood function (formulated in stimulus space). The uncertainty introduced by sensory noise, however, is transformed back through the inverse of the mapping function F (Eq. (3.8), Methods) between sensory and stimulus space; the very reason the likelihood function is asymmetric in the first place. b) Increasing the (symmetric) noise at the level of the sensory representation leads to a more asymmetric likelihood function (formulated in the stimulus space) and thus increases likelihood repulsion. As a result, the increase in prior attraction due to the increase in likelihood width is smaller than the increase in likelihood repulsion, leading to an overall net increase in repulsive bias. c) In contrast, adding (symmetric) stimulus noise does not affect the asymmetry of the likelihood function because the added noise essentially convolves the likelihood function with the noise kernel. The likelihood repulsion remains the same while the prior attraction grows because the overall width of the likelihood increases. As a result, the perceptual bias becomes more attractive (arrows). d) Summary plot illustrating how perceptual biases depend on stimulus and sensory noise. We assumed additive Gaussian noise and a squared-error loss function. Dots correspond to the conditions shown in b). In general, the perceptual bias is repulsive and grows with increasing sensory noise. However, increasing stimulus noise reduces the repulsive bias eventually leading to attractive biases for large noise levels. Note that this differential dependency on the different noise sources is a direct consequence of the inhomogeneous sensory representation imposed by Efficient coding. For comparison, the black curve illustrates the expected biases for a Bayesian observer model that simply assumes a symmetric likelihood function.

likelihood function, the increase in likelihood repulsion generally dominates, leading to a net increase in repulsive bias (Fig. 3.3b). Experimentally, we assume that sensory noise (or rather the signal-to-noise ratio) can be modulated by changing stimulus contrast or presentation time.

In summary, our Bayesian observer model predicts that perception is often biased away from the peak of the prior. Furthermore, it predicts that internal and external noise can differentially modulate these biases: increasing internal noise increases repulsive bias while increasing stimulus noise decreases repulsive bias, eventually leading to attractive perceptual biases. These predictions are surprising and at odds with predictions of standard Bayesian observer models.

3.2.3 Model validation against human psychophysical data

We validated the model predictions against measured perceptual biases for two visual stimulus variables with known natural stimulus distributions, local orientation θ and spatial frequency ξ.

Orientation perception

Several studies have measured the distribution of visual orientations in natural environments by carefully analyzing natural image data [37, 83]. The extracted distributions are fairly robust with regard to the specifics of the analysis and the image content (*e.g.*; indoor *versus* outdoor scenes [171]). These studies consistently

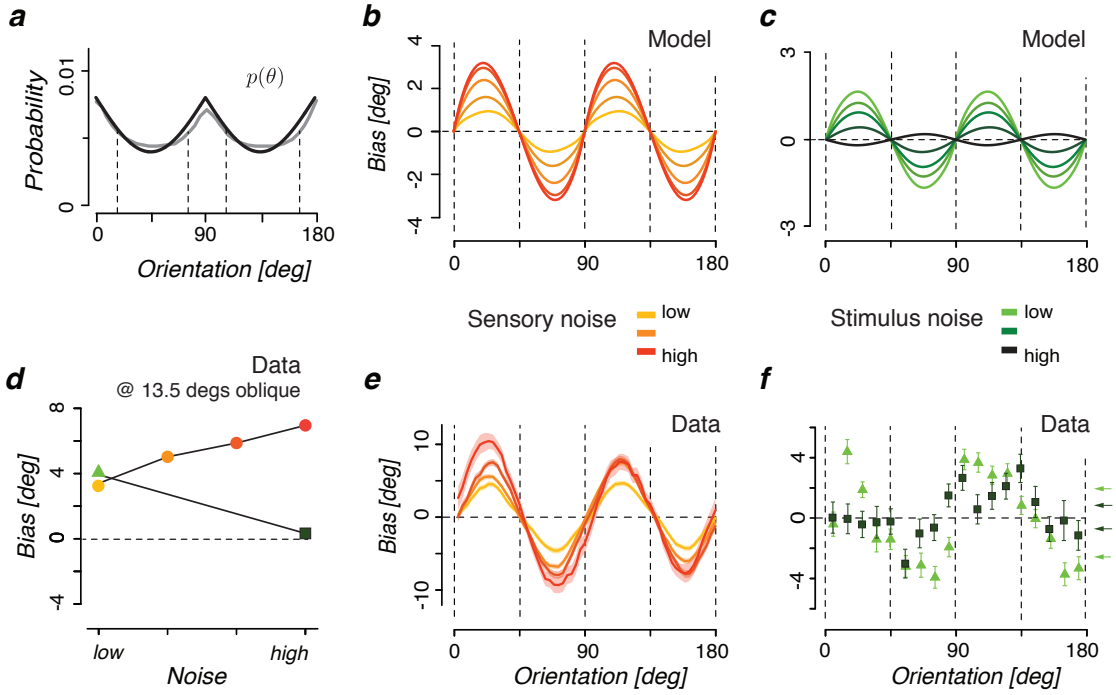

Figure 3.4: *Biases in perceived orientation.* a) Measured distribution of local visual orientation in natural images (gray line - re-plotted from [83]), superimposed with the parametric description used for the model predictions (black line: $p(\theta) = c_0(2 - |\sin\theta|)$ where c_0 is a normalization constant). b) Predicted mean biases as a function of stimulus orientation θ and different levels of sensory noise; biases are generally repulsive, *i.e.*, away from the nearest cardinal orientation, with larger biases for larger noise magnitudes. c) Same but for different levels of stimulus noise; here the repulsive biases are smaller for larger noise magnitudes, eventually becoming attractive. Curves in b,c) represent the expected bias values over the full measurement distributions. d) Measured biases at 15 degrees oblique orientations (average over all four orientations indicated by dashed lines in a)). Data from [175, 50]. The biases well match the predicted behavior shown in Fig. 3.3d. e) Measured biases as a function of sensory noise ($\pm$1 SEM). Sensory noise was modulated by different stimulus presentation times (low to high: 1000ms, 160ms, 80ms, 40ms). Reanalyzed data from [50]. f) Measured biases for two levels of additive Gaussian stimulus noise. Arrows indicate the mean bias over all orientations within each of two corresponding quadrants (*e.g.* black top arrow: mean bias for high stimulus noise computed over the range (0,45)$\cup$(90,135) degs). The overall biases are clearly repulsive and are reduced for larger stimulus noise. Replotted from [175].

reported multimodal distributions with peaks at each of the two cardinal orientations (*i.e.*, horizontal and vertical). We used a parametric approximation of the measured distribution by Girshick and colleagues [83] in order to generate model predictions of perceived visual orientation (Fig. 3.4a - black line). Figures 3.4b,c show the predicted mean biases as a function of stimulus orientation θ for different levels of sensory and stimulus noise, respectively. The predicted biases are typically repulsive and thus toward the nearest oblique orientation. Biases are zero for the cardinal and oblique orientations yet reach their maximum for orientations that lie in between. These oblique biases have been reported as early as in the late 19th century [95], and several studies have supported these findings since [175, 50]. The shape of the bias curves as a function of stimulus orientation is similar for both noise types. However, we predict that the bias amplitude grows with increasing sensory noise (Fig. 3.4b) while it decreases for increasing stimulus noise and eventually flips its sign; *i.e.*, turns into an attractive bias (Fig. 3.4c). Psychophysical data from two recent studies support our predictions [175, 50]. Figures 3.4d,e,f show the measured perceptual bias for stimulus orientations at 15 degrees oblique as well as as for the entire range of orientations as a function of stimulus and sensory noise. The observed bias patterns well match our predictions shown in Fig. 3.3c, and Fig. 3.4b and c, respectively.

Note that two previous studies have proposed Bayesian observer models for the perception of visual orientation [175, 83]. The models were validated against

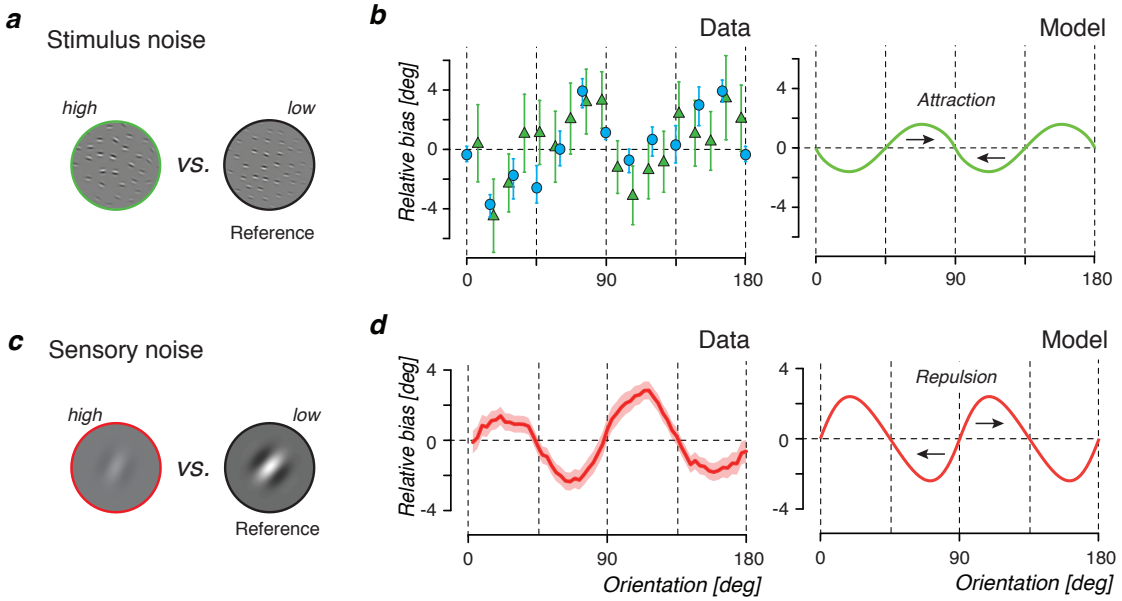

Figure 3.5: *Relative biases in perceived orientation.* Relative bias is the difference in perceived orientation between a high noise compared to a low noise stimulus (reference). a) Two orientation stimuli with different levels of stimulus noise. Each stimulus consists of an array of Gabor elements, and the width of the distribution from which the orientation of the elements are sampled controls the noise level. b) Measured relative biases as a function of stimulus orientation using the stimuli shown in a); Data replotted from [83] (blue) and [175] (green). The relative bias is attractive because the repulsive bias is smaller for the high noise stimulus (see Fig. 3.4f). c) Two orientation stimuli associated with different levels of sensory noise. Sensory noise can be modulated by stimulus contrast (this example) or presentation time (as in [50]) with lower contrast/shorter presentation time corresponding to higher sensory noise. d) Measured relative bias (± 1 SEM) between the percepts of two stimuli with different sensory noise as a function of stimulus orientation [50]. Relative bias is repulsive because the repulsive bias is larger for larger sensory noise. Our model accounts for both relative bias patterns while the previously proposed models by Tomassini *et al.* [175] and Girshick *et al.* [83] can only predict the relative bias for different stimulus noise (b) .

psychophysical measurements of *relative bias* between two stimuli with different levels of stimulus (external) noise. Specifically, both studies used the type of array stimuli shown in Fig. 3.5a and measured the difference in perceived orientation between a stimulus with high versus low stimulus noise. Although the percept of each of the two stimuli is biased toward the oblique orientations, it is less repulsive for the high noise stimulus (Fig. 3.4f). Thus the relative bias is indeed attractive and therefore can be accounted for by these models (Fig. 3.5b). However, they cannot explain the repulsive biases and their differential noise dependencies shown in Fig. 3.4, nor can they account for the relative bias between a stimulus with high versus low sensory noise (Fig. 3.5c). This relative bias is again repulsive because high sensory noise leads to larger repulsive biases (Fig. 3.4e). Thus, we predict that if Girshick and colleagues had fit their Bayesian model to 2AFC data collected with stimuli of different sensory rather than different stimulus noise, their fit prior distribution would not have matched the natural stimulus distribution (Fig. 3.4a) and would show peaks at the oblique orientations instead [83]. The results here suggest that the notion that perceived orientation is biased towards the cardinal axes because of a prior belief that favors cardinal orientations is simplistic.

Spatial frequency perception

We assume that the distribution of spatial frequencies ξ in natural visual environments is well represented by the empirically measured amplitude spectrum of natu-

ral images. Multiple studies reported spectra that approximately follow a power-law function $p(\xi) \propto 1/f^{\alpha}$ with values for α around one [147, 148] (Fig. 3.6a). We chose $\alpha = 1$ for simplicity but verified that our results are robust with regard to other values within the reported range. Because $p(\xi)$ is monotonically decreasing, we predict that in the absence of stimulus noise, perceived spatial frequency is biased toward higher frequencies across the entire frequency range. We also predict that increasing sensory noise (by *e.g.*, reducing stimulus contrast) biases the percept toward even higher frequency values (Fig. 3.6b) while increasing stimulus noise leads to a decrease in repulsive bias that eventually can turn into an attractive bias (Fig. 3.6c). Our predictions are consistent with psychophysically measured biases in perceived spatial frequency as a function of stimulus contrast [79] (Fig. 3.6d). Biases for different levels of stimulus noise have not been reported yet but could be measured using synthesized stimuli with different spectral bandwidths (see *e.g.* [137]).

Specifying the loss function

The proposed Bayesian observer model is fully specified for known natural stimulus distributions, with the exception of the loss function. The loss function is an integral part of any optimal Bayesian observer model, specifying how costly some perceptual errors are for the observer. The assumption is that the observer chooses an estimate (percept) that minimizes the expected loss (see Methods for details). Unfortunately, it is difficult to determine the actual loss function of a human observers when

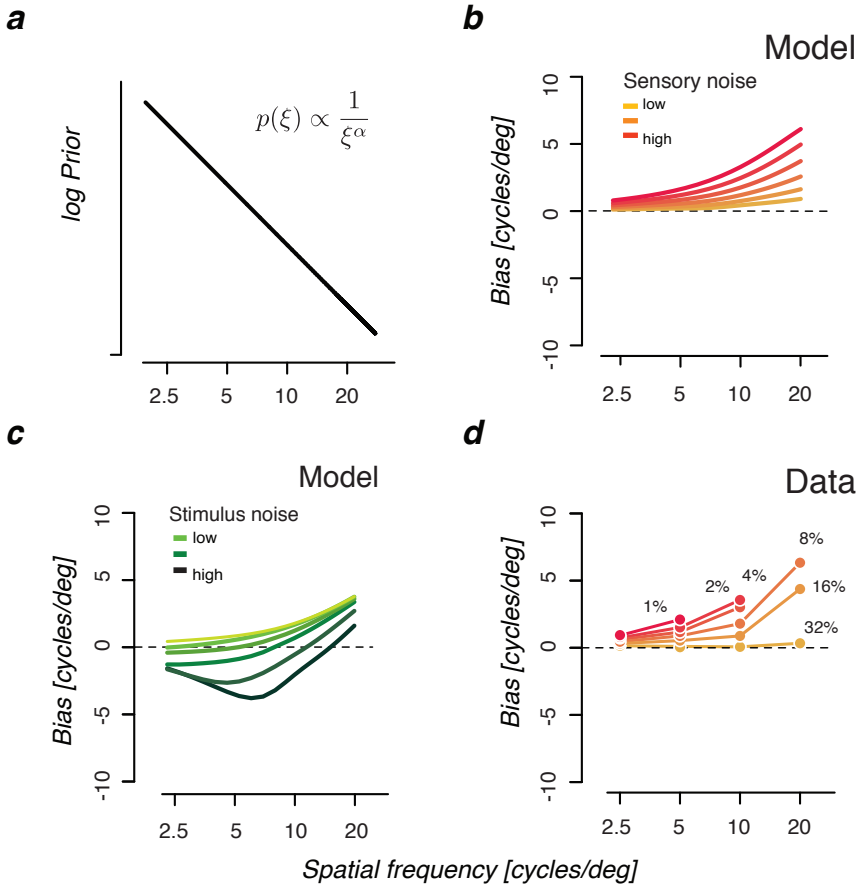

Figure 3.6: *Biases in perceived spatial frequency.* a) The distribution of spatial frequency in natural images approximately follows a power-law function of the form $p(\xi) \propto 1/f^{\alpha}$ with reported values for α around one [147]. Here we set $\alpha = 1$. b) The predicted biases as a function of spatial frequency for different levels of sensory (internal) noise. c) Predicted biases for different levels of stimulus (external) noise. d) Biases in perceived spatial frequency measured for different levels of sensory noise. Data replotted from [79]. The experiments used different levels of stimulus contrast (1, 2, 4, 8, 16, and 32%) to modulate sensory noise. Stimuli consisted of a Gabor patch with different spatial frequency. The predicted biases for stimulus noise in c) have not been validated yet. Note that at very low and very high spatial frequencies, the prior distribution is no longer well described by a single power-law function [147]. As a result, our predictions here are limited to the intermediate frequency-range, for which the prior is well approximated by a single power-law.

performing low-level perceptual tasks. Our predictions so far made the common assumption of a squared-error loss function (or L_2 norm), which is equivalent to computing the posterior mean. In order to explore the degree to which the model predictions depend on the specific choice of the loss function we compared them to predictions based on two other, also widely used loss functions from the L_p family: the L_0 loss (equivalent to the posterior mode, *i.e.*, a maximum a posteriori estimator) and the L_1 loss (equivalent to the posterior median, *i.e.*, a more robust estimator).

Figure 3.7 shows the predicted biases for the individual loss functions for both orientation θ and spatial frequency ξ and under conditions of both sensory and stimulus noise. Overall, the predictions for the the L_1 and L_2 loss are qualitatively similar although the bias magnitudes are smaller for the L_1 loss (Fig. 3.7a,c). This is expected since the median of the posterior is less repulsed than the mean. The reduced repulsive effect of the likelihood asymmetry also shows in the case of added stimulus noise (Fig. 3.7b,d). The transition from repulsive to attractive bias occurs at lower levels of stimulus noise. Predictions for the L_0 loss (MAP estimate), however, are distinctly different in that the bias is always attractive in these examples. The L_0 loss is unique in the sense that it does not take into account the shape of the posterior distribution. It is considered "degenerate" because it does not employ the information contained in the full posterior distribution. This intuitively explains why the predicted bias according to the MAP estimate is attractive: the repulsive

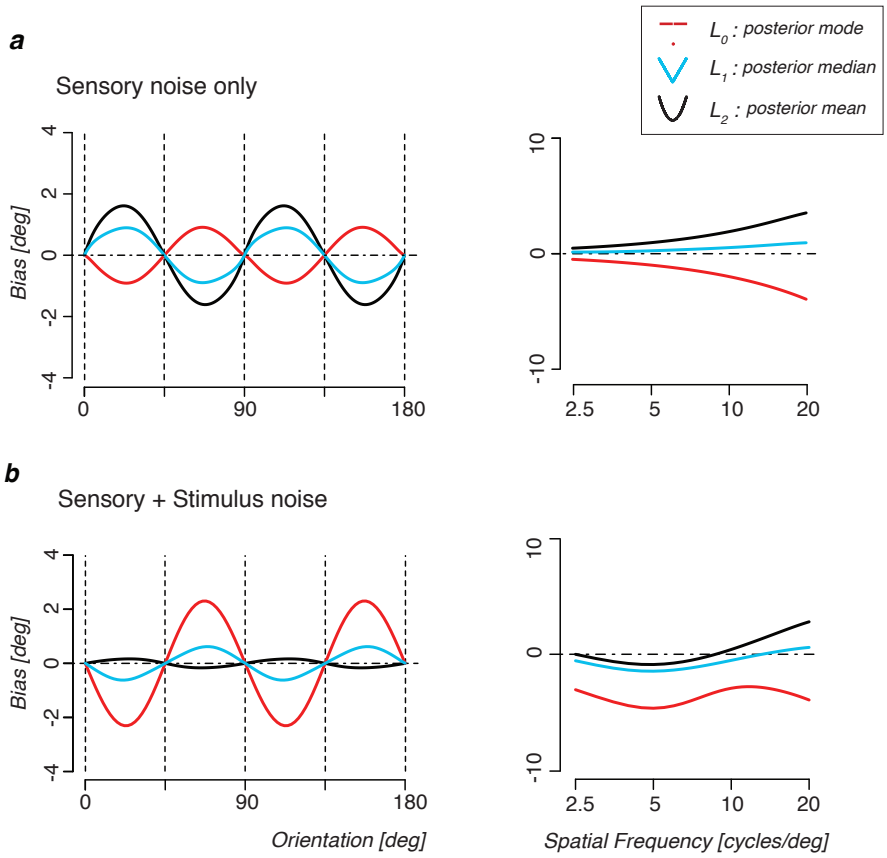

Figure 3.7: *Predicted biases for different loss functions.* a) Predicted biases in perceived orientation for the observer model with L_0-norm (posterior mode (MAP) - red), L_1-norm (posterior median - blue), or L_2-norm (posterior mean - black) loss function. Both the L_1- and L_2-norm predict repulsive biases while the L_0-norm always leads to attractive biases. Sensory noise is fixed and identical for all three models. b) Adding stimulus noise reduces the likelihood asymmetry and thus increases the attractive influence of the prior. The influence of the likelihood asymmetry is weaker with the L_1 loss compared to the L_2 loss, explaining the earlier transition to attractive biases with increasing stimulus noise magnitude. c,d) The same pattern is predicted for the perceptual biases in spatial frequency.

influence of the likelihood asymmetry is masked by the particular shape of the loss function and thus the prior attraction dominates. Any symmetric loss function other than the degenerate L_0 loss, however, preserves the repulsive influence of the likelihood asymmetry on the percept. Thus, the qualitative predictions of our observer model are fairly robust with regard to the specific choice of the loss function. The comparison also suggests that humans' perception of low-level stimuli is not guided by a degenerate loss function, which supports previous findings [105, 96].

3.3 Discussion

We have investigated the idea that the natural stimulus statistics not only determine how sensory information is represented, but also how this representation is interpreted to form a percept. Specifically, we introduced a new Bayesian observer model that is constrained by Efficient coding. As a result, the likelihood function and prior belief of our model are linked and jointly constrained by the natural stimulus statistics. The observer model makes two surprising and, at first sight, counter-intuitive general predictions. It predicts that under many conditions perceptual biases are repulsive *i.e.*, away from the peak of the prior distribution. It also predicts that sensory (internal) and stimulus (external) noise differentially affect perceptual bias when the stimulus distributions are non-homogeneous. We demonstrated that these predictions are confirmed by reported perceptual biases for visual orientation and spatial frequency, two perceptual variables for which the natural stimulus distri-

butions are known. The model accounts for biases measured over a wide range of noise and experimental conditions. In particular, the model provides a theoretical explanation for repulsive biases that previously proposed Bayesian observer models have failed to account for.

Our formulation of the Bayesian observer model is based on certain assumptions. For example, we considered a particular Efficient coding scheme (maximizing mutual information) although other formulations are also possible such as *e.g.*, minimizing redundancy [10] or reconstruction error [181], or formulations that take into account the coding requirements of downstream (motor) representations and actions [149]. In general, the choice of the encoding criterion depends on many constraints, not least on the task for which the encoded sensory information is used for [155]. For low-level stimulus variables (such as local visual orientation and spatial frequency) that are likely to be the basis for many different and more complex representations and tasks, optimizing for a more generic information criterion may represent a good encoding strategy of the visual system [10, 111]. As we show, the chosen formulation seems well supported by the data, yet it may need to be adjusted when modeling more cognitive stimulus variables or tasks.

Similarly, while we only considered scalar (*i.e.*, single-valued, one-dimensional) stimulus variables there is behavioral evidence that the brain can rapidly learn to efficiently encode also more complex stimulus variables (*e.g.* sound frequency spectra). Efficient coding solutions for these more complex variables, however,

may also differ from the solution presented here. Because different Efficient coding schemes impose different constraints on the shape of the likelihood, this may lead to different predictions for perceptual biases in all these cases. Such avenues would be interesting to explore in the future, although potential model predictions might be difficult to validate experimentally.

Although not explicitly specified in our model formulation we implicitly assumed a (quasi-)stationary perceptual environment and thus stationary stimulus distributions for our predictions. This assumption is probably valid for low-level stimulus variables (such as *e.g.*, spatial frequency or visual orientation), yet is certainly invalid under conditions where stationarity is explicitly violated such as in experiments that require subjects to rapidly learn a particular stimulus distribution (*e.g.*, [96, 32]), or during instances of perceptual adaptation. We have previously proposed that the characteristic repulsive adaptation aftereffects can be explained by asymmetric likelihood functions that result from an efficient re-distribution of sensory resources according to changes in the recent stimulus history [170]. The here proposed observer model uses a mathematically more rigid formulation and, in addition, imposes a tight link between prior belief, likelihood function, and stimulus distribution. It will be interesting to test to what degree our proposed observer model can account for adaptation aftereffects when formulated for stimulus distributions over shorter time-scales.

Our model formulation does not specify how the sensory measurement was ex-

tracted from low level sensory signals such *e.g.*, generating a measurement of local visual orientation based on the high-dimensional retinal image signal. Understanding this feature extraction process is important in characterizing what form of uncertainty and ambiguity is induced simply by the fact that under natural conditions stimulus variables such as *e.g.*, visual orientation are only indirectly encoded in the sensory signal [78]. Here, we focused on simple stimulus noise models that are sufficient to capture the typical noise characteristics of the artificial stimulus displays typically used in psychophysical experiments (see Fig. 3.4). However, there is no principled reason why the framework could not be extended to incorporate more complex uncertainty structures.

It is worth considering the implications for a potential physiological instantiation of the proposed perceptual encoding-decoding process. We purposefully used a formulation of Efficient coding (Eq. (3.1)) that is not based on detailed assumptions about the tuning characteristics of the underlying neural representation of the sensory information. This has the advantage that the formulation is sufficiently specific to define the likelihood function and thus allow clear predictions of perceptual behavior, yet is general in that it is not tied to any particular neural implementation (in contrast to *e.g.* our initial formulation [184]). We believe that is the right level of abstraction because it provides a fairly general observer model of perceptual behavior that can be widely applied. However, many equivalent neural encoding solutions are possible for a stimulus variable with a given distribution (prior). We demon-

strate this by considering three neural populations (Fig. 3.8b,c,d). Each population consists of neurons with independent Poisson firing statistics yet with quite different tuning characteristics in terms of neural density, tuning curve shape and response gain. Nonetheless, all three populations constitute equivalent efficient representations of the same stimulus variable θ with a distribution shown in Fig. 3.8a. The likelihood functions computed for each population's response are therefore similar (assuming that the gain is sufficiently large such that our assumption about noise symmetry is met) and show the expected asymmetries. As a result, Bayesian decoding of each of the three neural population leads to similar, repulsive bias curves (Fig. 3.8h,i,j).

Physiological constraints are also likely to influence the specific Efficient coding solution. For example, wiring constraints could limit the amount by which tuning curve widths can vary in a population, which would favor the solutions shown in Fig. 3.8b,d over the solution shown in Fig. 3.8c for a highly non-uniform stimulus distribution. Interestingly, this may provide an explanation for some of the differences in tuning characteristics between neurons in area V1 encoding orientation and neurons in area MST encoding heading direction, respectively. Perceptually, both stimulus variables exhibit similar repulsive biases away from the cardinal orientations [175, 50] (see Fig. 3.4), respectively from heading directions straight ahead or backwards [42, 45]. The measured neural tuning characteristics, however, are quite different: While the orientation tuning density and widths of neurons in V1

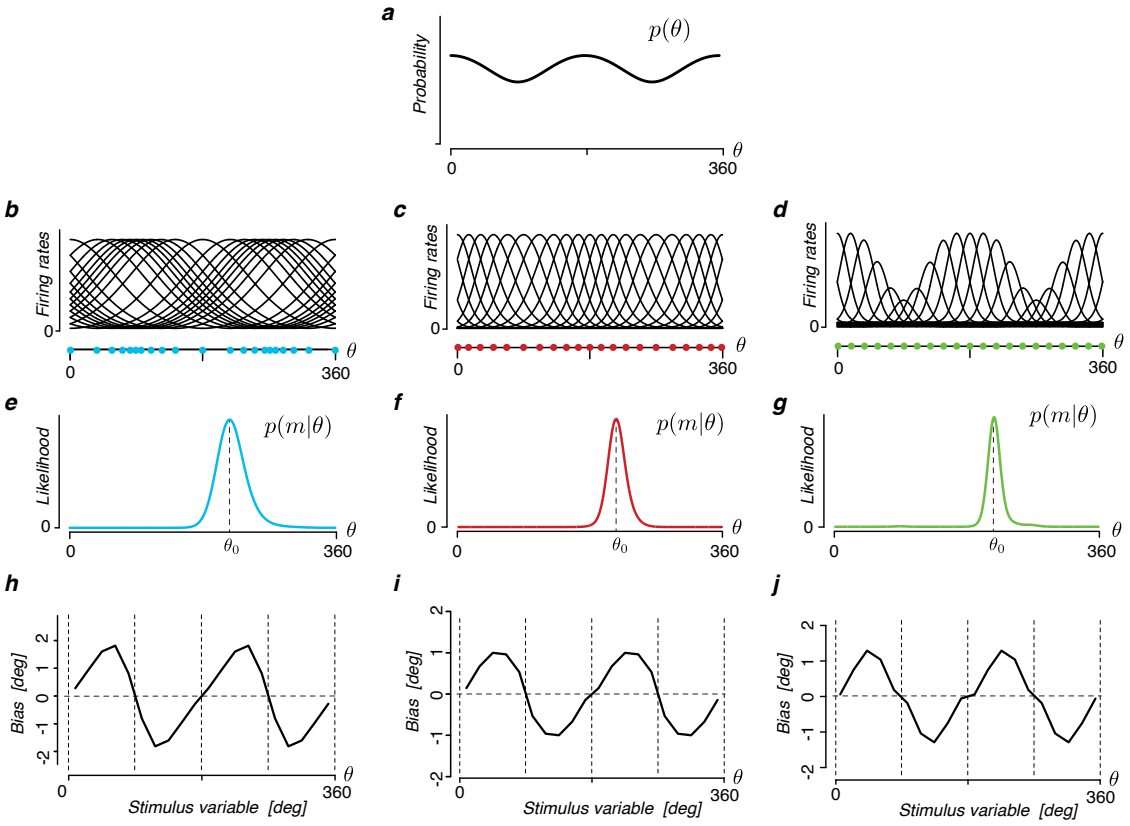

Figure 3.8: *Equivalent efficient neural representations for the same stimulus distribution.* a) A circular stimulus variable θ with stimulus distribution $p(\theta)$. Three different neural populations that efficiently encode θ according to Eq. (3.1): b) The tuning curves of the first population were constrained to be wide and identical (in stimulus space). Efficient coding promotes a distribution of the neurons such that the flanks of the individual tuning curves most overlap at the peaks of the prior distribution. The neural density (blue dots) is such that it is *lowest at the prior peaks*. c) The tuning curves as well as the density of a second population were allowed to vary [76, 77]. As a result, the neural density follows the prior distribution (red dots) [65] and tuning curves are narrowest at the prior peaks. d) The tuning curve shapes as well as the neural density (green dots) of the third population were constrained to be identical/homogeneous. Only the gain was allowed to vary. As a result, neurons at the peak of the prior had highest gain. e,f,g) Population likelihoods for all three populations (averaged over 400 presentations of the same stimulus value θ_0) are similar (up to a scale factor) and show the predicted asymmetry with the heavy tail pointing away from the nearest peak of the prior. h,i,j) As a result, Bayesian decoding of all three population results in similar repulsive biases. Biases are computed over 10000 samples of the neural population response.

are loosely in agreement with the population shown in Fig. 3.8c [91, 146], neurons in area MST rather resemble the population shown in Fig. 3.8b with the majority of neurons preferentially tuned to left- and right-ward directions [85]. Our findings suggest that both the population of V1 and MST neurons efficiently represent a stimulus variable with similar natural distributions, leading to similar perceptual biases, yet may be subject to additional constraints at the level of implementation. However, we currently do not have a good estimate of the natural stimulus distribution for heading direction, which would allow us to confirm this conjecture.

Several neural implementations of Bayesian inference have been proposed, which use decoding mechanisms that are similar to the population vector read-out [65, 66, 183, 77]. The implementations all rely on neural populations whose tuning densities match the prior distribution. Interestingly, note that the population shown in Fig. 3.8c has these tuning characteristics and could be readily decoded with such a population vector read-out, thus providing a neural implementation of our observer model. Whether other, equivalent Efficient encoding solutions (see *e.g.* Fig. 3.8b,d) also allow for simple and physiologically plausible decoding mechanisms is an interesting question to explore in future research.

An obvious question is how our proposed Bayesian observer model and its predictions are in agreement with the results of previous studies that show the characteristic "biases toward the prior" behavior. First, it is important to note that our observer model does not exclusively predict repulsive biases. For example, as

stimulus noise gets large biases become attractive. The same applies when considering stimuli that are in a range where the prior is not monotonic over the support of the likelihood (*e.g.* stimuli close to the peak of a unimodal prior distribution). Also, measured percepts depend on the specifics of the experimental setup, and thus what looks like an attractive bias might be *e.g.*, a relative difference between repulsive biases (see Fig. 3.5). Finally, some previous results may have relied on incorrect assumptions about the stimulus distribution, again, with the result that biases that appear to be attractive may actually be repulsive. The formulation of our new Bayesian observer model is general and we think it will allow us to explain perceptual biases far beyond those examples shown in this paper, including biases that currently cannot be explained. The problem we see for such future investigations is to obtain good estimates of the relevant stimulus distributions, which is often difficult (*e.g.*, see [57] for distributions of visual speed). But even if this information is not available or too difficult to obtain the proposed observer model is better constrained, allowing improved fits to psychophysical data with fewer free parameters compared to previous Bayesian modeling approaches [169, 83].

Last but not least, our work addresses the common criticism that Bayesian observer models are not well constrained and thus can explain *post hoc* essentially any data with the appropriate choice of prior belief and likelihood function [99, 21]. We have shared this concern to some degree as we have expressed in the past [169]. However, we think we have addressed this criticism by introducing a

better constrained Bayesian observer model that, at the same time, also can explain perceptual data that were previously unaccounted for. We think that Bayesian models with arbitrarily chosen parametric descriptions have served their purpose, providing an intuitive understanding of how prior beliefs may affect perception. While the focus on prior beliefs was important, our results demonstrate that it can lead to a rather simplistic understanding of the Bayesian modeling approach, which also fails to capture various interesting aspects of perceptual behavior (such as the repulsive biases). Our new observer model is a next step in elaborating the Bayesian hypothesis, putting the focus on a more principled definition of the likelihood function and the way different noise sources affect perceptual processing.

Methods

Efficient encoding We assumed an Efficient coding constraint that maximizes the mutual information between a scalar stimulus variable θ and its sensory representation m [121, 10]. Fisher information $J(\theta)$, defined as

$$J(\theta) = \int (\frac{\partial \ln p(m|\theta)}{\partial \theta})^2 p(m|\theta) dm \ , \tag{3.2}$$

can be used to specify a bound on mutual information in the asymptotic limit [27]. Assuming the bound is tight, mutual information can be expressed as

$$I[\theta, m] = \frac{1}{2}\ln(\frac{S^2}{2\pi e}) - KL(p(\theta)||\frac{\sqrt{J(\theta)}}{S}), \tag{3.3}$$

where $S = \int_\theta \sqrt{J(\theta)}d\theta$ ([127]). S can be intuitively understood as the total coding resource available. $KL(\ ||\ )$ represents the Kullback-Leibler (KL) divergence [110], which is always non-negative.

The goal is to choose $J(\theta)$ to maximize $I[\theta, m]$ for a fixed $p(\theta)$. Technically, this requires us to impose an additional constraint on Fisher information. We require the total Fisher information to be bounded, *i.e.*,

$$S = \int_\theta \sqrt{J(\theta)}d\theta \leq C\ . \tag{3.4}$$

For any value of C, maximizing mutual information requires the above KL divergence term to be zero. This is equivalent to

$$p(\theta) \propto \sqrt{J(\theta)}\ . \tag{3.5}$$

This relationship has been previously derived under slightly different assumptions [27, 127, 76].

Bayesian decoding Bayesian decoding consists of defining an estimate $\hat{\theta}(m)$ of the stimulus value given measurement m such that the expected loss according to a loss function $L(\hat{\theta}(m), \theta)$ is minimal, *i.e.*,

$$\text{argmin}_{\hat{\theta}} \int p(\theta|m) L(\hat{\theta}(m), \theta) d\theta. \tag{3.6}$$

The key quantity here is $p(\theta|m)$, which represents the posterior probability distribution over θ for a given sensory measurement m. According to Bayes' rule [12], the posterior can be computed as $p(\theta|m) \propto p(\theta)p(m|\theta)$, where $p(\theta)$ is the prior belief and $p(m|\theta)$ represents the likelihood function on θ. For the specific L_0, L_1, and L_2 loss functions considered in this paper, the optimal estimator $\hat{\theta}(m)$ is the posterior mode, median, and mean, respectively.

Perceptual bias for L_2 loss (sensory noise only) We can analytically derive the bias $b(\theta)$ of our Bayesian observer model in the case of a squared error loss function (L_2 loss) assuming no stimulus noise. The posterior mean can be computed in terms of the likelihood functions and the prior belief as following

$$\hat{\theta}_{L_2}(m) = \frac{\int \theta p(m|\theta) p(\theta) d\theta}{\int p(m|\theta) p(\theta) d\theta} \ . \tag{3.7}$$

With the Efficient coding assumption above Eq. (3.5) we can now express the bias as a function of the prior belief. First, we define a one-to-one mapping $F(\theta)$

that transforms the stimulus space to a *sensory space* with units $\tilde{\theta} = F(\theta)$ for which the Fisher Information (as well as the stimulus distribution) is uniform [111, 127]. The mapping is defined as

$$F(\theta) = \int_{-\infty}^{\theta} p(\chi) d\chi \ , \tag{3.8}$$

which is the cumulative of the prior distribution $p(\theta)$.

We then re-write the estimate Eq. (3.7) by replacing θ with the inverse of the mapping, *i.e.*, $\theta = F^{-1}(\tilde{\theta})$. Given a sensory measurement m, we can write the estimator as

$$\hat{\theta}_{L_2}(m) = \frac{\int F^{-1}(\tilde{\theta}) p(m|F^{-1}(\tilde{\theta})) p(F^{-1}(\tilde{\theta})) dF^{-1}(\tilde{\theta})}{\int p(m|F^{-1}(\tilde{\theta})) p(F^{-1}(\tilde{\theta})) dF^{-1}(\tilde{\theta})} = \frac{\int F^{-1}(\tilde{\theta}) p(m|F^{-1}(\tilde{\theta})) d\tilde{\theta}}{\int p(m|F^{-1}(\tilde{\theta})) d\tilde{\theta}} \ . \tag{3.9}$$

With

$$K(m, \tilde{\theta}) = \frac{p(m|F^{-1}(\tilde{\theta}))}{\int p(m|F^{-1}(\tilde{\theta})) d\tilde{\theta}} \tag{3.10}$$

we can further simplify the notation and get

$$\hat{\theta}_{L_2}(m) = \int F^{-1}(\tilde{\theta}) K(m, \tilde{\theta}) d\tilde{\theta} \ . \tag{3.11}$$

In order to get the expected value of the estimate, $\langle \hat{\theta}_{L_2} \rangle$, for a particular stimulus

value θ_0 we marginalize Eq. (3.11) over the measurement space M for θ_0, thus

$$\begin{aligned}\langle\hat{\theta}_{L_2}\rangle_{\theta_0} &= \int\int_M p(m|\theta_0)F^{-1}(\tilde{\theta})K(m,\tilde{\theta})dmd\tilde{\theta} \\ &= \int F^{-1}(\tilde{\theta})\int_M p(m|\theta_0)K(m,\tilde{\theta})dmd\tilde{\theta} \\ &= \int F^{-1}(\tilde{\theta})L_{\theta_0}(\tilde{\theta})d\tilde{\theta}\ ,\end{aligned} \tag{3.12}$$

where we define

$$L_{\theta_0}(\tilde{\theta}) = \int_M p(m|\theta_0)K(m,\tilde{\theta})dm\ . \tag{3.13}$$

Therefore, $L_{\theta_0}(\tilde{\theta})$ is the expected normalized likelihood function expressed in the sensory space given a particular stimulus value θ_0. We assume that $L_{\theta_0}(\tilde{\theta})$ is symmetric around the true stimulus value $\tilde{\theta}_0$ in this space. Thus, with Eq. (3.11) we then can compute the expected bias at θ_0 as

$$b(\theta_0) = \int F^{-1}(\tilde{\theta})L_{\theta_0}(\tilde{\theta})d\tilde{\theta} - F^{-1}(\tilde{\theta}_0) \tag{3.14}$$

Assuming the prior density to be smooth, we expand F^{-1} in the neighborhood $(\tilde{\theta_0}-h,\tilde{\theta_0}+h)$, which covers the support of the likelihood function. Using a first-order Taylor expansion with mean-value form of the remainder, we get

$$F^{-1}(\tilde{\theta}) = F^{-1}(\tilde{\theta_0}) + F^{-1}(\tilde{\theta_0})'(\tilde{\theta}-\tilde{\theta}_0) + \frac{1}{2}F^{-1}(\tilde{\theta}_x)''(\tilde{\theta}-\tilde{\theta}_0)^2\ , \tag{3.15}$$

with $\tilde{\theta}_x$ guaranteed to exist in between $\tilde{\theta}_0$ and $\tilde{\theta}$. By re-writing Eq. (3.14) in terms of this expansion, we find that

$$\begin{aligned} b(\theta_0) &= \int_{\tilde{\theta}_0-h}^{\tilde{\theta}_0+h} \frac{1}{2} F^{-1}(\tilde{\theta}_x)''_{\tilde{\theta}} (\tilde{\theta}-\tilde{\theta}_0)^2 L_{\theta_0}(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{2} \int_{\tilde{\theta}_0-h}^{\tilde{\theta}_0+h} (\frac{1}{p(F^{-1}(\tilde{\theta}_x))})'_{\tilde{\theta}} (\tilde{\theta}-\tilde{\theta}_0)^2 L_{\theta_0}(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{2} \int_{\tilde{\theta}_0-h}^{\tilde{\theta}_0+h} -(\frac{p(\theta_x)'_\theta}{p(\theta_x)^3})(\tilde{\theta}-\tilde{\theta}_0)^2 L_{\theta_0}(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{4} \int_{\tilde{\theta}_0-h}^{\tilde{\theta}_0+h} (\frac{1}{p(\theta_x)^2})'_\theta (\tilde{\theta}-\tilde{\theta}_0)^2 L_{\theta_0}(\tilde{\theta}) d\tilde{\theta}. \end{aligned} \tag{3.16}$$

In general, there is no simple rule to judge the sign of $b(\theta_0)$, because θ_x varies with θ and the sign of $(1/p(\theta_x)^2)'$ thus may change. However, if the prior is monotonic on the interval $F^{-1}((\tilde{\theta}_0-h, \tilde{\theta}_0+h))$ then the sign of $(\frac{1}{p(\theta_x)^2})'$ is always the same as the sign of $(\frac{1}{p(\theta_0)^2})'$ and therefore, the sign of $b(\theta_0)$ is the same as the sign of $(\frac{1}{p(\theta_0)^2})'$. This means that the bias and the local slope of the prior have opposite signs. It implies that the bias is repulsive, *i.e.*, *away* from the peak of the prior.

Additionally, in the small noise regime where the likelihood is sufficiently narrow, the prior can always be approximated as being monotonic over the support of the likelihood function. Due to the continuity of $(\frac{1}{p(\theta)^2})'$, we can approximate $(\frac{1}{p(\theta_x)^2})'$ by $(\frac{1}{p(\theta_0)^2})'$ and thus write the bias as

$$b(\theta_0) \approx C(\frac{1}{p(\theta_0)^2})' \,, \tag{3.17}$$

where C is a positive constant.

The key assumption we made in the above derivation is that the average likelihood function $L_{\theta_0}(\tilde{\theta})$ in the sensory space $(\tilde{\theta})$ is symmetric. The dimensionality of the measurement m is not important, *i.e.*, m can be a scalar or a vector (*e.g.* response vector of a neural population), as long as the assumption that $L_{\theta_0}(\tilde{\theta})$ is symmetric is approximately true.

Perceptual bias under more general conditions Under more general conditions that include stimulus noise and/or different loss functions, the expected perceptual bias can no longer be computed analytically. However, numerical solutions can be computed for general conditions according to the encoding-decoding cascade description of the proposed Bayesian observer model. In particular, we can distinguish the effect of stimulus versus sensory noise (Fig. 3.4 and 3.6) by modeling the sensory measurement m as

$$m = F(\theta + \delta_s) + \delta_n \ , \tag{3.18}$$

where δ_s represents the stimulus noise (expressed in stimulus space) and δ_n the sensory noise (expressed in sensory space). We assumed the sensory noise to be Gaussian (respectively, vonMises) distributed, and the stimulus noise to follow the actual noise distributions used in the psychophysical experiments we modeled (often Gaussian/vonMises distributed as well). The transformation F that imposes the

Efficient coding constraint determines how the stimulus noise is mapped to the sensory space (Eq. (3.8)). For any stimulus value θ_0 the conditional probability $p(m|\theta_0)$ can be computed according to Eq. (3.18) and the specific noise distributions. For each m, we can numerically compute the Bayesian estimator $\hat{\theta}(m)$ according to a specific loss function (L_0, L_1, L_2) using Eq. (3.6). Finally, for any given stimulus value θ_0, the expected bias $b(\theta_0)$ can be computed by marginalizing the estimate $\hat{\theta}(m)$ over the measurement distribution $p(m|\theta_0)$ and then subtracting the true value θ_0, thus

$$b(\theta_0) = \int \hat{\theta}(m) p(m|\theta_0) dm - \theta_0 \ . \tag{3.19}$$

Neural simulation We applied a little trick in order to generate three neural populations that have different tuning characteristics yet match in their Fisher information $J(\theta)$ (up to a scaling factor) and satisfy the efficiency constraint Eq. (3.5). We first generated the population in Fig. 3.8b by assuming that it consists of $N = 20$ neurons with wide and uniform tuning curves (vonMises distribution) whose preferred tuning follow an arbitrary density distribution $d(\theta) \propto 1.2 - |\cos\theta|$. We then computed the population Fisher information assuming independent Poisson noise, and with Eq. (3.5) derived the stimulus distribution (*i.e.*, the prior belief) $p(\theta)$ (Fig. 3.8a). The tuning curves of the second population (Fig. 3.8c) were obtained by re-parameterizing a set of homogeneous tuning curves through the cumulative prior $F(\theta)$ as previously proposed [76, 77]. To create the third neural population in (Fig. 3.8d), we started from a homogeneous set of tuning curves with relatively

narrow tuning widths and adjusted the gain of individual neurons such that the square-root of the population Fisher information matched the prior distribution. Numerically, this is done via a non-negative least squares fit. These procedures guarantees that all three populations have identical Fisher information (up to a scale factor) and thus are efficient representations of the prior distribution. The likelihoods shown in Fig. 3.8e,f,g represent the average likelihoods computed over 400 samples of the population responses for a fixed stimulus value θ_0. The biases (Fig. 3.8h,i,j) are computed by drawing 10000 samples assuming independent Poisson-spiking neuron models, and calculating the average bias of the Bayes' least squares estimator over the samples while exploiting the symmetry in the stimulus distribution.

Data re-analysis The bias curves shown in Fig. 3.4e were obtained by re-analyzing the data set presented by DeGardelle and colleagues [50]. In their experiments, stimulus orientation was randomly sampled over the entire range (*i.e.*, [0, 180] degs). Bias was computed by averaging the trials over a sliding window (3 degs size). The resulting bias $b(\theta)$ was then further smoothed with a boxcar filter with width w = 45 degs. We performed this analysis for four stimulus conditions corresponding to stimulus presentation times of 40, 80, 160, and 1000ms. For these conditions, the shape and amplitudes of the bias curves were robust with regard to the chosen bin size and the width of the smoothing kernel. A fifth stimulus condition corresponding to a presentation time of 20ms was excluded in our analysis because the amplitude

of the bias curve was dependent on the bin size, making it impossible to reliably determine the magnitude of the bias. The error bars for individual orientations θ_0 in Fig. 3.4e represent the circular standard error, which was estimated based on the data samples within the window $[\theta_0 \pm 22.5]$ degs. Relative bias shown in Fig. 3.5d was calculated by taking the difference between the biases corresponding to the 160ms and 1000ms stimulus presentation conditions reported in Fig. 3.4e. The error bar in Fig. 3.5d was calculated as the square root of the sum of the squared SEM in Fig. 3.4e (160ms and 1000ms conditions).

Chapter 4

The sense of place: grid cells in the brain and the transcendental number e

4.1 Introduction

How does the brain represent space? Tolman[174] suggested that the brain must have an explicit neural representation of physical space, a *cognitive map*, that supports higher brain functions such as navigation and path planning. The discovery of place cells in the rat hippocampus [132, 133] suggested one potential locus for this map. Place cells have spatially localized firing fields which reorganize dramatically when the environment changes [117]. Another potential locus for the cognitive map

of space has been uncovered in the main input to hippocampus, a structure known as the medial entorhinal cortex (MEC) [75, 86]. When rats freely explore a two dimensional open environment, individual "grid cells" in the MEC display spatial firing fields that form a periodic triangular grid which tiles space (Fig. 4.1A). It is believed that grid fields provide relatively rigid coordinates on space based partly on self-motion and partly on environmental cues [129]. The scale of grid fields varies systematically along the dorso-ventral axis of the MEC (Fig. 4.1A)[86]. Recently it was shown that grid cells are organized in discrete modules within which the cells share the same orientation and periodicity, but vary randomly in phase [86, 164].

How does the grid system represent spatial location and what function does the modular variation in grid scale serve? Here, we propose that the grid system provides a hierarchical representation of space where fine grids provide precise location and coarse grids resolve ambiguity, and that the grids are organized to minimize the number of neurons required to achieve the behaviorally necessary spatial resolution. Consistent with studies of grid cell and place cell remapping, our analyses assume that there is a behaviorally defined maximum range over which a fixed grid represents locations[73]. Our hypotheses, together with general assumptions about tuning curve shape and decoding mechanism, give a rationale for the triangular lattice structure of two-dimensional grid cell firing maps and predict a geometric progression of grid scales. Crucially, the theory further predicts that the ratio of adjacent grid scales will be modestly variable within and between animals with a

mean in the range $1.4 - 1.7$ depending on the assumed decoding mechanism used by the brain. With additional assumptions the theory also predicts that the ratio between grid scale and individual grid field widths should lie in the same range. These predictions naturally explain the structural parameters of grid cell modules measured in rodents [11, 164, 81]. Our results follow from general principles, and thus we expect similar organization of the grid system in other species. The theory makes further predictions including: (a) the number of grid scales necessary to support navigation over typical behavioral distances, (b) deficits in spatial behavior that will obtain upon inactivating specific grid modules, and (c) the structure of one and three dimensional grids that will be relevant to navigation in, e.g., bats. Remarkably, in a simple decoding scheme, the scale ratio in an n-dimensional environment is predicted to be close to $\sqrt[n]{e}$.

As we will explain, our results, and their apparent experimental confirmation in [164], suggest that the grid system implements a two-dimensional neural analog of a base-b number system. This provides an intuitive and powerful metaphor for interpreting the representation of space in the entorhinal cortex.

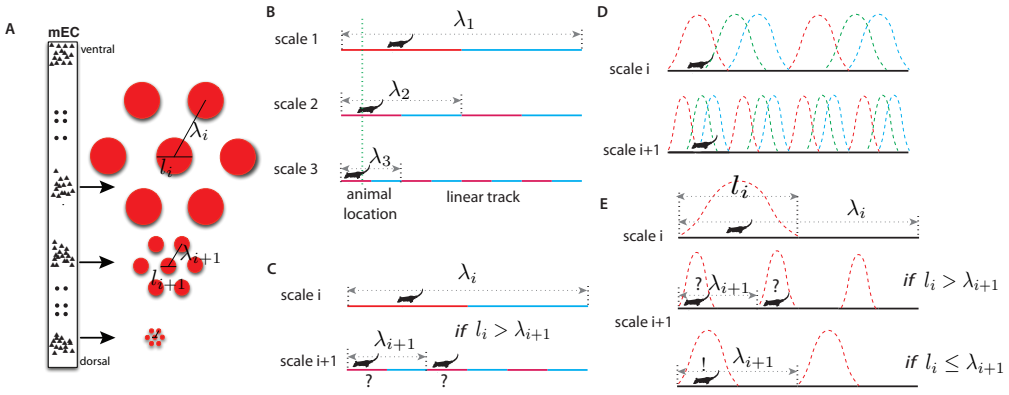

Figure 4.1: Representing place in the grid system. **(A)** Grid cells (small triangles) in the medial entorhinal cortex (MEC) respond when the animal is in a triangular lattice of physical locations (red circles) [75, 86]. The scale of periodicity (the "grid scale", λ_i) and the size of the regions evoking a response above a noise threshold (the "grid field width", l_i) vary modularly along the dorso-ventral axis of the MEC [86]. Grid cells within a module vary in the phase of their spatial response, but share the same period and grid orientation (in two dimensions) [164]. **(B)** A simplified binary grid scheme for encoding location along a linear track. At each scale (λ_i) there are two grid cells (red vs. blue firing fields). The periodicity and grid field widths are halved at each successive scale. **(C)** The binary scheme in (B) is ambiguous if the grid field width at scale i exceeds the grid periodicity at scale $i+1$. E.g., if the grid fields marked in red respond at scales i and $i+1$, the animal might be in either of the two marked locations. **(D)** The grid system is composed of discrete modules, each of which contains neurons with periodic tuning curves, and varying phase, in space. **(E)** For a simple Winner-Take-All decoder (see main text) of the grids in panel D, decoded position will be ambiguous unless $l_i \leq \lambda_{i+1}$, analogously to panel C (see text). Variants of this limitation occur in other decoding schemes.

4.2 Results

4.2.1 The setup

The key features of the grid system in the MEC are schematized in Fig. 4.1A. Grid cells are organized in modules, and cells within a module share a common lattice organization of their firing fields [11, 164]. These lattices have periods $\lambda_1 > \lambda_2 > \cdots \lambda_m$, measured as the distance between nearest neighbor firing fields. It will prove convenient to define "scale factors" $r_i = \lambda_i/\lambda_{i+1}$ relating the periods of adjacent scales. In each module, the grid firing fields (i.e. the connected spatial regions that evoke firing) are compact (with a diameter denoted l_i) after thresholding for activity above the noise level (see, e.g., [86]). Within any module, grid cells have a variety of spatial phases so that at least one cell will respond at any physical location (Fig. 4.1B,D). Grid modules with smaller field widths l_i provide more local spatial information than those with larger scales. However, this increased spatial precision comes at a cost: the correspondingly smaller periodicity λ_i of these modules leads to increased ambiguity since there are more grid periods within a given spatial region (e.g., see scale 3 in the schematic one dimensional grid in Fig. 4.1B,D. By contrast, modules with large periods and field widths have less spatial precision, but also less ambiguity (e.g. in scale 1 in Fig. 4.1B the red cell has only one firing field in the environment and hence no ambiguity).

We propose that the Entorhinal cortex exploits this tradeoff to implement a

hierarchical representation of space where large scales resolve ambiguity and small scales provide precision. Consistently with existing data for one and two dimensional grids [26, 11, 164], we will take the largest grid period λ_1 to be comparable to the range over which space is represented unambiguously by a fixed grid without remapping [73]. (An alternative view, that the range might greatly exceed the largest period, is addressed in the Discussion.) The spatial resolution of such a grid can be measured by comparing the range of spatial representation set by the largest period λ_1 to the precision (related to the smallest grid field width l_m) to quantify how many distinct spatial "bins" can be resolved. We will assume that the required resolution is set by the animal's behavioral requirements.

4.2.2 Intuitions from a simplified model

What are the advantages of a multi-scale, hierarchical representation of physical location? Consider an animal living in an 8m linear track and requiring spatial precision of 1m to support its behavior. To develop intuition, consider a simple model where location is represented in the animal's brain by reliable neurons with rectangular firing fields (e.g., Fig. 4.1B). The animal could achieve the required resolution in a *place coding* scheme by having eight neurons tuned to respond when the animal is in 1m wide, non-overlapping regions. Consider an alternative, the idealized *grid coding* scheme in Fig. 4.1B. Here the two neurons at the largest scale (λ_1) have 4m wide tuning curves so that their responses just indicate the left and

right halves of the track. The pairs of neurons at the next two scales have grid field widths of 2m and 1m respectively and proportionally shorter periodicities as well. These pairs successively localize the animal into 2m and 1m bins. All told only 6 neurons are required, less than in the place coding scheme. This suggests that grid schemes that integrate multiple scales of representation can encode space more efficiently, i.e. with fewer neural resources. In the sensory periphery there is evidence of selection for more efficient circuit architectures (e.g. [155]). If similar selection operates in cortex, the experimentally measured grid architecture should be predicted by maximizing the efficiency of the grid system given a behaviorally determined range and resolution. Thus we seek to predict the key structural parameters of the grid system – the ratios $r_i = \lambda_i/\lambda_{i+1}$ relating adjacent scales (which need not be equal).

The need to avoid spatial ambiguity constrains the ratios r_i. Again in our simple model, consider Fig. 4.1C where the cells with the grid fields marked in red respond at scales i and $i+1$. Then the animal might be in either of the two marked locations. Avoiding ambiguity requires that λ_{i+1}, the period at scale $i+1$, must exceed l_i, the grid field width at scale i. Variants of this condition will recur in the more realistic models that we will consider. Theoretically, one could resolve the ambiguity in Fig. 4.1C by combining the responses of more grid modules, provided they have mutually incommensurate periods [64, 162]. However, anatomical and functional evidence suggests that place cells selectively read out contiguous subsets

of Entorhinal grid modules along the dorso-ventral axis [180, 160]. For each of these restricted readouts to provide a well-defined spatial map, ambiguities like the one in Fig. 4.1C should be resolved at each scale. The hierarchical position encoding schemes that we consider below embody this observation by seeking to reduce position ambiguity at each scale, given the responses at larger scales.

4.2.3 Efficient grid coding in one dimension

How should the grid system be organized to minimize the resources required to represent location unambiguously with a given resolution? Consider a one dimensional grid system that develops when an animal runs on a linear track. As described above, the ith module is characterized by a period λ_i while the ratio of adjacent periods is $r_i = \lambda_i/\lambda_{i+1}$. Within any module, grid cells have periodic, bumpy response fields with variety of spatial phases so that at least one cell responds at any physical location (Fig. 4.1D). If d cells respond above the noise threshold at each point, the number of grid cells n_i in module i will be $n_i = d\lambda_i/l_i$. We will take d, the *coverage factor* to be the same in each module. In terms of these parameters, the total number of grid cells is $N = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} d\frac{\lambda_i}{l_i}$ where m is the number of grid modules. How should such a grid be organized to minimize the number of grid cells required to achieve a given spatial resolution? The answer might depend on how the brain decodes the grid system. Hence we will consider decoding methods at extremes of decoding complexity, and show that they give similar answers for

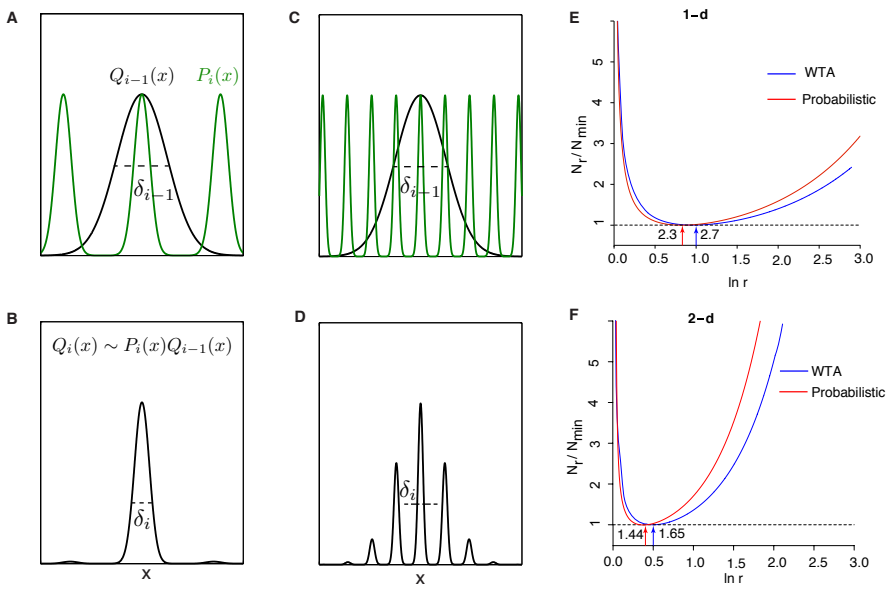

Figure 4.2: **(A-D)** Trade-off between precision and ambiguity in the probabilistic decoder. **(A)** The probability of position x given the responses of all grid cells at scales larger than module i is described by the distribution $Q_{i-1}(x)$ (black curve), and the uncertainty in position is given by the standard deviation δ_{i-1}. The probability of position given just the responses in module i will be a periodic function $P_i(x)$ (green curve). **(B)** The probability distribution over position x after combining module i with all larger scales is $Q_i(x) \sim P_i(x)Q_{i-1}(x)$, and has reduced uncertainty δ_i. **(C)** Precision can be improved by increasing the scale factor, thereby narrowing the peaks of $P_i(x)$. However, the periodicity shrinks as well, increasing ambiguity. **(D)** The distribution over position $Q_i(x)$ from combining the modules shown in *C*. Ambiguity from the secondary peaks leads to an overall uncertainty δ_i larger than in *B*, despite the improved precision from the narrower central peak. **(E)** The optimal ratio r between adjacent scales in a hierarchical grid system in one dimension for a simple Winner-Take-All decoding model (blue curve, WTA) and a probabilistic decoder (red curve). Here N_r is the number of neurons required to represent space with resolution R given a scaling ratio r, and $N_{\min}$ is the number of neurons required at the optimum. In both decoding models, the ratio $N_r/N_{\min}$ is independent of resolution, R. For the Winner-Take-All model, $N_r \propto r/\ln r$, while the curve for the probabilistic model is derived numerically (mathematical details in Supplementary Information). The Winner-Take-All model predicts $r = e \approx 2.7$, while the probabilistic decoder predicts $r \approx 2.3$. The minima of the two curves lie within each others' shallow basins. **(F)** Same as *E*, but in two dimensions with a triangular grid. The minima occur at $r = \sqrt{e} \approx 1.65$ for Winner-Take-All and $r \approx 1.4$ for the probabilistic case (mathematical details in Supplementary Information). The shallowness of the basins around these minima predicts that some variability of adjacent scale ratios is tolerable, both within and between animals.

the optimal grid.

First imagine a decoder which considers the animal as localized within the grid fields of the most responsive cell in each module [38, 124]. A simple "winner-take-all" (WTA) scheme of this kind can be easily implemented by neural circuits where lateral inhibition causes the influence of the most responsive cell to dominate. A maximally conservative decoder ignoring all information from other cells and from the shape of the tuning curve (illustrated in Fig. 4.1E) could then take uncertainty in spatial location to be equal to l_i. The smallest interval that can be resolved in this way will be l_m. We therefore quantify the resolution of the grid system (the number of spatial bins that can be resolved) as the ratio of the largest to the smallest scale, $R_1 = \lambda_1/l_m$, which we assume to be large and fixed by the animal's behavior. In terms of scale factors $r_i = \lambda_i/\lambda_{i+1}$, we can write the resolution as $R_1 = \prod_{i=1}^{m} r_i$, where we also defined $r_m = \lambda_m/l_m$. As in our simplified model above, unambiguous decoding requires that $l_i \leq \lambda_{i+1}$ (Fig. 4.1C,E), or, equivalently, $\frac{\lambda_i}{l_i} \geq r_i$. To minimize $N = d\sum_i \lambda_i/l_i$, all the $\frac{\lambda_i}{l_i}$ should be as small as possible; so this fixes $\frac{\lambda_i}{l_i} = r_i$. Thus we are reduced to minimizing the sum $N = d\sum_{i=1}^{m} r_i$ over the parameters r_i, while fixing the product $R_1 = \prod_i r_i$. Because this problem is symmetric under permutation of the indices i, the optimal r_i turn out to all be equal, allowing us to set $r_i = r$ (Supplementary Material, Sec. 2). This is our first prediction: **(1)** the ratios between adjacent periods will be constant. The constraint on resolution then gives $m = \log_r R$, so that we seek to minimize $N(r) = d\, r \log_r R_1$

with respect to r: the solution is $r = e$ (Fig. 4.2E; Supplementary Material). This gives a second prediction: **(2)** the ratio of adjacent grid periods should be close to $r = e$. Therefore, for each scale i, $\lambda_i = e\,\lambda_{i+1}$ and $\lambda_i = e\,l_i$. This gives a third prediction: **(3)** the ratio of the grid period and the grid field width will be constant across modules, and be close to the scale ratio.

More generally, in winner-take-all decoding schemes, the local uncertainty in the animal's location in grid module i will be proportional to the grid field width l_i. The proportionality constant will will be a function $f(d)$ of the coverage factor d that depends on the tuning curve shape. Thus, the uncertainty will be $f(d)\,l_i$. Unambiguous decoding at each scale requires that $\lambda_{i+1} \geq f(d) l_i$. The smallest interval that can be resolved in this way will be $f(d)\,l_m$, and this sets the positional accuracy of the decoding scheme. Finally we require that $\lambda_1 > L$ where L is a scale big enough to ensure that the grid code resolves positions over a sufficiently large range. Behavioral requirements fix the required positional accuracy and range. The optimal grid satisfying these constraints is derived in the Supplementary Information. Again, the adjacent modules are organized in a geometric progression and the ratio between adjacent periods is predicted to be e. However, the ratio between the grid period and grid field width in each module depends on the specific model through the function $f(d)$. Thus, within winner-take-all decoding schemes, the constancy of the scale ratio, the value of the scale ratio, and the constancy of the ratio of grid period to field width are parameter-free predictions, and therefore furnish tests

of theory. If the tests succeed, $f(d)$ can be matched to data to constrain possible mechanisms used by the brain to decode the grid system.

What do we predict for a general decoding scheme that optimally pools all the information available in the responses of cells within and between modules? Statistically, the best we can do is to use all these responses, which may individually be noisy, to find a probability distribution over physical locations that can then inform subsequent behavioral decisions. Thus the population response at each scale i gives rise to a likelihood function over location $P(x|i)$, which will have the same periodicity λ_i as the individual grid cells' firing rates (Fig. 4.2A). This likelihood explicitly captures the uncertainty in location given the tuning and noise characteristics of the neural population in the module i. Because there are scores of neurons in each grid module [164], $P(x|i)$ can approximated as a periodic sum of Gaussians without making restrictive assumptions about the shapes of the tuning curves of individual grid cells. The standard deviations of the peaks in $P(x|i)$, which we call σ_i, depend on the tuning curve shape and response noise of individual grid cells, and will decrease as the coverage factor d increases. To have even coverage of space, the number of grid phases, and thus grid cells in a module, must be uniformly distributed so that equally reliable posterior distributions can be formed at each point in the unit cell of the module response.

This requires that the number of cells (and phases) in the module should be proportional to the ratio $\frac{\lambda_i}{\sigma_i}$. Summing over modules, the total number of grid cells

will be $N \propto \sum_{i=1}^{m} \frac{\lambda_i}{\sigma_i}$. The composite posterior given all m scales and a uniform prior over positions, $Q_m(x)$, will be given by the product $Q_m(x) \propto \Pi_{i=1}^{m} P(x|i)$, assuming independent response noise across scales (Fig. 4.2B). The animal's overall uncertainty about its position depends on the standard deviation δ_m of the composite posterior distribution $Q_m(x)$. Setting δ_0 to be the uncertainty in location without using any grid responses at all, we can quantify resolution as $R = \delta_0/\delta_m$.

In this framework, there is a precision-ambiguity tradeoff controlled by the scale factors r_i. The larger these ratios, the more rapidly grid field widths shrink in successive modules, thus increasing precision and reducing the number of modules, and hence grid cells, required to achieve a given resolution. However, if the periods of adjacent scales shrink too quickly, the composite posterior $Q_i(x)$ will develop prominent side-lobes (Fig. 4.2C,D) making decoding ambiguous as reflected in a large standard deviation δ_i of the composite posterior distribution (Fig. 4.2B,D). This ambiguity could be avoided by shrinking the width of $Q_{i-1}(x)$ – however, this would require increasing the number of neurons $n_1, \cdots n_{i-1}$ in the modules $1 \cdots i-1$. Ambiguity can also be avoided by having a smaller scale ratio (so that the side lobes of the posterior $P(x|i)$ of module i do not penetrate the central lobe of the composite posterior $Q_{i-1}(x)$ of modules $1 \cdots i-1$. But reducing the the scale ratios to reduce ambiguity increases the number of modules necessary to achieve the required resolution, and hence increases the number of grid cells. This sets up a tradeoff – increasing the scale ratios reduces the number of modules to achieve

a fixed resolution, but requires more neurons in each module; reducing the scale ratios permits the use of fewer grid cells in each module, but increases the number of required modules. Optimizing this tradeoff (analytical and numerical details in Supplementary Information; Fig. 2E) predicts: **(1)** A constant scale ratio between the periods of each grid module, and **(2)** An optimal ratio $r \approx 2.3$, slightly smaller than, but close to the winner-take-all value, e. The theory also predicts a fixed ratio between grid period λ_i and posterior likelihood width σ_i. However, the relationship between σ_i and the more readily measurable grid field width l_i depends on a variety of parameters including the tuning curve shape, noise level and neuron density.

Why is the predicted scale factor based on the probabilistic decoder somewhat smaller than the prediction based on the winner-take-all analysis? In the probabilistic analysis, when the likelihood is combined across modules, there will be side lobes arising from the periodic peaks of the likelihood derived from module i multiplying the tails of the Gaussian arising from the previous modules. These side lobes increase location ambiguity (measured by the standard deviation δ_i of the overall likelihood). Reducing the scale factor reduces the height of side lobes because the secondary peaks from module i move further into the tails of the Gaussian derived from the previous modules. Thus, conceptually, the optimal probabilistic scale factor is smaller than the winner-take-all case in order to suppress side lobes that arise in the combined likelihood across modules (Fig. 4.2). Such side lobes were absent in the winner-take-all analysis. The theory also predicts a fixed ratio between grid pe-

riod λ_i and posterior likelihood width σ_i. However, the relationship between σ_i and the more readily measurable grid field width l_i depends on a variety of parameters including the tuning curve shape, noise level and neuron density.

The minima for both the probabilistic decoder and the winner-take-all decoder are shallow (Fig. 4.2E), so that the scaling ratio r may lie anywhere within a basin around the optimum at the cost of a small number of additional neurons. Even though our two decoding strategies lie at extremes of complexity (one relying just on the most active cell at each scale and another optimally pooling information in the grid population) their respective "optimal intervals" substantially overlap. That these two very different models make overlapping predictions suggests that our theory is robust to variations in the detailed shape of grid cells' grid fields and the precise decoding model used to read their responses. Moreover, such considerations also suggest that these coding schemes have the capacity to tolerate developmental noise: different animals could develop grid systems with slightly different scaling ratios, without suffering a large loss in efficiency.

4.2.4 General grid coding in two dimensions

How do these results extend to two dimensions? Let λ_i be the distance between nearest neighbor peaks of grid fields of width l_i (Fig. 4.1A). Assume in addition that a given cell responds on a lattice whose vertices are located at the points $\lambda_i(n\mathbf{u} + m\mathbf{v})$, where n, m are integers and $\mathbf{u}, \mathbf{v}$ are linearly independent vectors generating

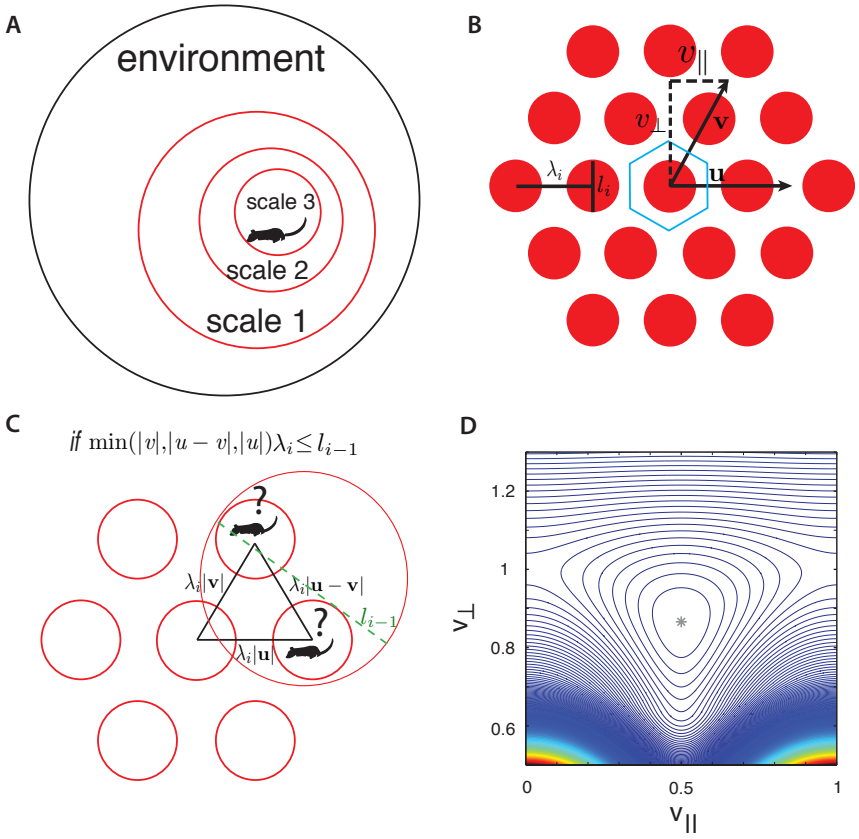

Figure 4.3: **(A)** Circular firing fields in a two dimensional grid scheme. **(B)** A general two-dimensional lattice is parameterized by two vectors $\mathbf{u}$ and $\mathbf{v}$ and a periodicity parameter λ_i. Take $\mathbf{u}$ to be a unit vector, so that the spacing between peaks along the $\mathbf{u}$ direction is λ_i, and denote the two components of $\mathbf{v}$ by $v_\parallel$, $v_\perp$. The blue-bordered region is a fundamental domain of the lattice, the largest spatial region that can be unambiguously represented. **(C)** The two dimensional analog of the ambiguity in Fig. 1C, E for the winner-take-all decoder. If the grid fields in scale i are too close to each other relative to the size of the grid field of scale $i-1$ (i.e. l_{i-1}), the animal might be in one of several locations. **(D)** Contour plot of normalized neuron number $N/N_{\min}$ in the probabilistic decoder, as a function of the grid geometry parameters $v_\perp, v_\parallel$ after minimizing over the scale factors for fixed resolution R. As in Fig. 2E,F, the normalized neuron number is independent of R. The spacing between contours is 0.01, and the asterisk labels the minimum at $v_\parallel = 1/2, v_\perp = \sqrt{3}/2$; this corresponds to the triangular lattice.

the lattice (Fig. 4.3B). We may take $\mathbf{u}$ to have unit length ($|\mathbf{u}| = 1$) without loss of generality, however $|\mathbf{v}| \neq 1$ in general. It will prove convenient to denote the components of $\mathbf{v}$ parallel and perpendicular to $\mathbf{u}$ by $v_{\|}$ and $v_{\perp}$, respectively (Fig. 4.3B). The two numbers $v_{\|}, v_{\perp}$ quantify the geometry of the grid and are additional parameters that we may optimize over: this is a primary difference from the one-dimensional case. We will assume that $v_{\|}$ and $v_{\perp}$ are independent of scale; this still allows for relative rotation between grids at different scales. At each scale, grid cells have different phases so that at least one cell responds at each physical location. The minimal number of phases required to cover space is computed by dividing the area of the unit cell of the grid ($\lambda_i^2||\mathbf{u} \times \mathbf{v}|| = \lambda_i^2|v_{\perp}|$) by the area of the grid field. As in the one-dimensional case, we define a coverage factor d as the number of neurons covering each point in space, giving for the total number of neurons $N = d|v_{\perp}| \sum_i (\lambda_i/l_i)^2$.

As before, consider a situation where grid fields thresholded for noise lie completely within compact regions and assume a simple decoder which selects the most activated cell and does not take tuning curve shape into account [38, 124, 49]. In such a model, each scale i simply serves to localize the animal within a circle of diameter l_i. The spatial resolution is summarized by the square of the ratio of the largest scale λ_1 to the smallest scale l_m: $R_2 = (\lambda_1/l_m)^2$. In terms of the scale factors $\tilde{r}_i = \lambda_i/\lambda_{i+1}$ we write $R_2 = \prod_{i=1}^{m} \tilde{r}_i^2$, where we also define $\tilde{r}_m = \lambda_m/l_m$. To decode the position of an animal unambiguously, each cell at scale $i+1$ should have at

most one grid field within a region of diameter l_i. We therefore require that the shortest lattice vector of the grid at scale i has a length greater than l_{i-1}, in order to avoid ambiguity (Fig. 4.3C). We wish to minimize N, which will be convenient to express as $N = d|v_\perp| \sum_i \tilde{r}_i^2 (\lambda_{i+1}/l_i)^2$. There are two kinds of contributions here to the number of neurons – the factors $\tilde{r}_i^2$ are constrained by the overall resolution of the grid, while, as we will see, the combination $|v_\perp|(\lambda_{i+1}/l_i)^2$ measures a packing density of discs placed on the grid lattice. This suggests that we separate the minimization of neuron number into first optimizing the lattice, and then optimizing ratios. After doing so, we can check that the result is the global optimum.

To obtain the optimal lattice geometry, we can ignore the resolution constraint, as it depends only on the scale factors and not the grid geometry. We may then exploit an equivalence between our optimization problem and the optimal circle-packing problem. To see this connection, consider placing disks of diameter l_i on each vertex of the grid at scale $i+1$. In order to avoid ambiguity, all points of the grid $i+1$ must be separated by at least l_i: equivalently, the disks must not overlap. The density of disks is proportional to $l_i^2/(\lambda_{i+1}^2|v_\perp|)$, which is proportional to the reciprocal of each term in N. Therefore, *minimizing* neuron number amounts to *maximizing* the packing density; and the no-ambiguity constraint requires that the disks do not overlap. This is the optimal circle packing problem, and its solution in two dimensions is known to be the triangular lattice [173], so $v_\parallel = 1/2$ and $v_\perp = \sqrt{3}/2$. Furthermore, the grid spacing should be as small as allowed by the

no-ambiguity constraint, giving $\lambda_{i+1} = l_i$.

We have now reduced the problem to minimizing $N = \frac{d\sqrt{3}}{2} \sum_i \tilde{r}_i^2$, over the scale factors $\tilde{r}_i$, while fixing the resolution R_2. This optimization problem is mathematically the same as in one dimension if we formally set $r_i \equiv \tilde{r}_i^2$. This gives the optimal ratio $\tilde{r}_i^2 = e$ for all i (Fig. 4.2F). We conclude that in two dimensions, the optimal ratio of neighboring grid periodicities is $\sqrt{e} \approx 1.65$ for the simple winner-take-all decoding model, and the optimal lattice is triangular.

The optimal probabilistic decoding model from above can also be extended to two dimensions with the posterior distributions $P(x|i)$ becoming sums of Gaussians with peaks on the two-dimensional lattice. In analogy with the one-dimensional case, we then derive a formula for the resolution $R_2 = \lambda_1/\delta_m$ in terms of the standard deviation δ_m of the posterior given all scales. δ_m may be explicitly calculated as a function of the scale factors $\tilde{r}_i$ and the geometric factors $v_\parallel, v_\perp$, and the minimization of neuron number may then be carried out numerically (Supplementary Material). In this approach the optimal scale factor turns out to be $\tilde{r}_i \approx 1.4$ (Fig. 4.2F), and the optimal lattice is again triangular (Fig. 4.3D). Attractor network models of grid formation readily produce triangular lattices [28]; our analysis suggests that this architecture is functionally beneficial in reducing the required number of neurons.

Once again, the optimal scale factors in both decoding approaches lie within overlapping shallow basins, indicating that our proposal is robust to variations in grid field shape and to the precise decoding algorithm (Fig. 4.2F). In two dimensions,

the required neuron number will be no more than 5% of the minimum if the scale factor is within the range $(1.43, 1.96)$ for the winner-take-all model and the range $(1.28, 1.66)$ for the probabilistic model. These "optimal intervals" are narrower than in the one-dimensional case, and have substantial overlap.

In summary, for 2-d case, the theory predicts that: (**1**) the ratios between adjacent scales should be a constant; (**2**) The optimal scaling constant is $\sqrt{e} \approx 1.65$ in a simple WTA decoding model, and it is $\approx 1.4.$ in a probabilistic decoding model; (**3**) The predictions off optimal grid field width depends on the specific decoding methods. (**4**) The grid lattice should be a triangular lattice.

4.2.5 Comparison to experiment

Our predictions agree with experiment [11, 81, 164] (see Supplementary Material for details of the data re-analysis). Specifically, Barry et al., 2007 (Fig. 4.4A) reported the grid periodicities measured at three locations along the dorso-ventral axis of of the MEC in rats and found ratios of ~ 1, ~ 1.7 and $\sim 2.5 \approx 1.6 \times 1.6$ relative to the smallest period[11] . The ratios of adjacent scales reported in [11] had a mean of 1.64 ± 0.09 (mean $\pm$ std. dev., $n = 6$), which almost precisely matches the mean scale factor of $\sqrt{e}$ predicted from the winner-take-all decoding model, and is also consistent with the probabilistic decoding model. In another study[108], the scale ratio between the two smaller grid scales, measured by the ratio between the grid frequencies, is reported to be ~ 1.57 in one animal. Recent analysis based

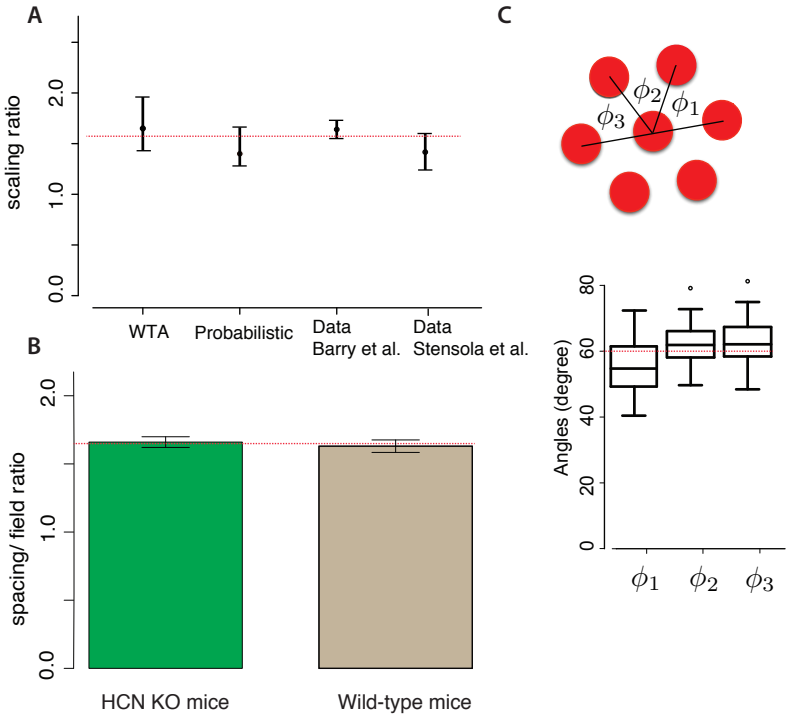

Figure 4.4: **(A)** Our models predict grid scaling ratios that are consistent with experiment. 'WTA' (Winner-Take-All) and 'probabilistic' represent predictions from two decoding models; the dot is the scaling ratio minimizing the number of neurons, and the bars represent the interval within which the neuron number will be no more than 5% higher than the minimum. For the experimental data, the dot represents the mean measured scale ratio and the error bars represent ± one standard deviation. Data were replotted from [11, 164]. The dashed red line shows a consensus value running through the two theoretical predictions and the two experimental datasets. **(B)** The mean ratio between grid periodicity (λ_i) and the diameter of grid fields (l_i) in mice (data from [81]). Error bars indicate ± one S.E.M. For both wild type mice and HCN knockouts (which have larger grid periodicities) the ratio is consistent with $\sqrt{e}$ (dashed red line). **(C)** The response lattice of grid cells in rats forms an equilateral triangular lattice with 60° angles between adjacent lattice edges (replotted from [86], $n = 45$ neurons from 6 rats). Dots represent outliers, as reported in [86].

on a larger data set [164] confirms the geometric progression of the grid scales in individual animals over four modules. The mean ratio between adjacent scales is 1.42 ± 0.17 (mean $\pm$ std. dev., $n = 24$) in that data set, accompanied by modest variability within and between animals. These measurements again match both our models (Fig. 4.4A). The optimal grid was triangular in both of our models, again matching measurements (Fig. 4.4C) [86, 129, 164].

A recent study measured the ratio between grid periodicity and grid field size to be 1.63 ± 0.035 (mean $\pm$ S.E.M., $n = 48$) in wild type mice[81]. This ratio was unchanged, 1.66 ± 0.03 (mean $\pm$ S.E.M., $n = 86$), in HCN1 knockout strains whose absolute grid periodicities increased relative to the wild type[81]. Such measurements are consistent with the prediction of the simple Winner-Take-All model, which predicts a ratio between grid period and grid field width of $\lambda_i/l_i = \sqrt{e} \approx 1.65$. (Fig. 4.4B).

4.3 Discussion

We have shown that a grid system with a discrete set of periodicities, as found in the Entorhinal cortex, should use a common scale factor r between modules to represent spatial location with the fewest neurons. In other words, the periods of grid modules should be organized in a geometric progression. In one dimension, this organization may be thought of intuitively as implementing a neural analog of a base-b number system. Roughly, the largest scale localizes the animal into a coarse region of the

environment and finer scales successively subdivide the region into b "bins". For example, suppose that the largest scale has one firing field in the environment and that $b = 2$, so that subsequent scales subdivide this firing field into halves (Fig. 4.1A,B). Then, keeping track of which half the animal occupies at each scale gives a binary encoding of location. This is just like a binary number system being used to encode a number representing the location. Our problem of minimizing neuron number while fixing resolution is analogous to minimizing the product of the number of digits and the number of decimal places (which we can term *complexity*) needed to represent a given range R of integers in a base-b number system. The complexity is approximately $C \sim b \log_b R$. What "base" minimizes the complexity of the representation? We can compute this by evaluating the extremum $\partial C/\partial b = 0$, and find that the optimum is at $b = e$ (details in Supplementary Material). Our full theory is a generalization of this simple fixed-base representational scheme for numbers to noisy neurons encoding two-dimensional location. It is remarkable that natural selection seems to have reached such efficient solutions for encoding location.

Our theory quantitatively predicted the ratios of adjacent scales within the variability tolerated by the models and by the data (Fig. 4.4). Further tests of our theory are possible. For example, a direct generalization of our reasoning says that in n-dimensions the optimal ratio between grid scales for winner-take-all decoding is $\sqrt[n]{e}$ (as compared to $\sqrt{e}$ in two dimensions). The three dimensional case is possibly relevant to the grid system in, e.g., bats [187, 186]. Robustly, for any given

decoding scheme, our theory would predict a smaller scaling ratio for 3d grids than for 2d grids. The packing density argument given above for two dimensional lattice structure, when generalized to three dimensions, would predict a face center cubic lattice. Bats are known to have 2d grids when crawling on surfaces [187] and if they also have a 3d grid system when flying, similar to their place cell system [186], our predictions for three dimensional grids can be directly tested. In general, the theory can be tested by comprehensive population recordings of grid cells along the dorsoventral axis for animals moving in one, two and three dimensional environments.

Our theory also predicts a logarithmic relationship between the natural behavioral range and the number of grid modules. To estimate the number of modules, m, required for a given resolution R_2 via the approximate relationship $m = \log R_2 / \log \tilde{r}^2$. Assuming that the animal must be able to represent an environment of area $\sim (10\,\mathrm{m})^2$ (e.g. [47]), with a positional accuracy on the scale of the rat's body size, $\sim (10\,\mathrm{cm})^2$, we get a resolution of $R_2 \sim 10^4$. Together with the predicted two-dimensional scale factor $\tilde{r}$, this gives $m \approx 10$ as an order-of-magnitude estimate. Indeed, in [164], 4-5 modules were discovered in recordings spanning up to 50% of the dorsoventral extent of MEC; extrapolation gives a total module number consistent with our estimate.

How many grid cells do we predict in total? Consider the simplest case where grid cells are independent encoders of position in two dimensions. Our likelihood analysis (see Sec. 3 of Supplementary Information for details) gives the number of

neurons as $N = mc(\lambda/\sigma)^2$ where m is the number of modules and c is constant. In detail, c is determined by factors like the tuning curve shape of individual neurons and their firing rates, but broadly what matters is the typical number of spikes K that a neuron emits during a sampling time, because this will control the precision with which location can be inferred from a single cell's response. General considerations [48] indicate that c will be proportional to $1/K$. We can estimate that if a rat runs at $\sim 50\,\mathrm{cm/s}$ and covers $\sim 1\,\mathrm{cm}$ in a sampling time, then a grid cell firing at $10\,Hz$ [164] gives $K \sim 1/5$. Using our prediction that the number of modules will be ~ 10 and that $\lambda/\sigma \approx 5.3$ in the optimal grid (see Supplementary Section 3), we get $N_{\mathrm{est}} \approx 1400$. This estimate assumed independent neurons and that the decoder of the grid system will efficiently use all the information in every grid cell's response. This is unlikely to be the case. Given homogeneous positive noise correlations within a grid module, which will arise naturally if grid cells are formed by an attractor mechanism, the required number of neurons could be an order of magnitude higher [161, 5]. Thus, in round numbers, we estimate that our theory requires something in the range of $\sim 1400 - 14000$ grid cells.

Are there so many grid cells in the Medial Entorhinal Cortex? In fact, we need this number of grid cells separately in Layer II and Layer III of the MEC since these regions likely maintain separate grid codes. (To see this, recall that Layers II and III project largely to the dentate gyrus and CA1 respectively [167, 56], while the place map in CA1 survives lesions of the dentate input to CA1 via CA3 [25].)

Physiological studies [150] have shown that only about 10% of the cells in MEC are Layer II grid cells, and another 10% are Layer III grid cells. Cells that have weak responsiveness during spatial tasks are probably undersampled in such experiments and so the real proportion of grid cells is likely to be somewhat smaller. Other studies [130] have shown that MEC has $\sim 10^5$ neurons. Thus we can estimate that Layer II and Layer III each contain some $5000 - 10000$ grid cells. This is well within the predicted theoretical range.

Our analysis assumed that grid code is hierarchical, with large grids resolving the spatial ambiguity created by the multiple firing fields of the small grids that deliver precision of location. Recall that place cells are thought to provide one readout of the grid system. Anatomical evidence [180] shows that the projections from the mEC to the hippocampus are restricted along the dorso-ventral axis, so that a given place cell receives input from perhaps a quarter of the mEC. The data of Stensola et al. [164] show additionally that the dorsal mEC is impoverished in large grid modules. Together with the anatomy [180], the hierarchical view of location coding that we have proposed then predicts that dorsal place cells should be revealed to have multiple place fields in large environments because their spatial ambiguities will not be fully resolved at larger scales. Preliminary evidence for this prediction has appeared in [62, 143].

We assumed that the largest scales of grid module should be roughly comparable to the behavioral range of the animal. This is consistent with the existing data

on grid modules [164] and with measurements in the largest environments tested so far [26] (periods at least as large as 10m in an 18m track). To accommodate very large environments, grids could either increase their scale (as reported at least transiently in [11, 164]) or could segment the environment into large sections [53, 52] across which remapping occurs [73]. These predictions can be tested in detail by exploring spatial coding in natural environments of behaviorally appropriate size and complexity. In fact, ethological studies have indicated a typical homing rate of a few tens of meters for rats with significant variation between strains [47, 68, 168, 156, 22] . Our theory predicts that the period of the largest grid module and the number of modules will be correlated with homing range.

In our theory, we took the coverage factor d (the number of grid fields overlapping a given point in space) to be the same for each module. In fact, experimental measurements have not yet established whether this parameter is constant, or varies between modules. How would a varying d affect our results? The answer depends on the dimension of the grid. In two dimensions, if neurons have weakly correlated noise, modular variation of the coverage factor does not affect the optimal grid at all. This is because the coverage factor cancels out of all relevant formulae, a coincidence of two dimensions (see Sec. 3 of the Supplementary Material, and p. 112 of [48]). In one and three dimensions, variation of d between modules will have an effect on the optimal ratios between the variable modules. Thus, if the coverage factor is found to vary between grid modules for animals navigating one

and three dimensions, our theory can be tested by comparing its predictions for the corresponding variations in grid scale factors. Similarly, even in two dimensions, if noise is correlated between grid cells, then variability in d can affect our predicted scale factor. This provides another avenue for testing our theory.

The simple winner-take-all model assuming compact grid fields predicted a ratio of field width to grid period that matched measurements in both wild-type and HCN1 knockout mice [81]. Since the predicted grid field width is model dependent, the match with the simple WTA prediction might be providing a hint concerning the method the brain uses to read the grid code. Additional data on this ratio parameter drawn from multiple grid modules may serve to distinguish and select between potential decoding models for the grid system. The probabilistic model did not make a direct prediction about grid field width; it instead worked with the standard deviation of the posterior $P(x \mid i)$, σ_i. This parameter is predicted to be $\sigma_i = 0.19\lambda_i$ in two dimensions (see Supplementary Material). This prediction could be tested behaviorally by comparing discrimination thresholds for location to the period of the smallest module. The standard deviation σ_i can also be related to the noise, neural density and tuning curve shape in each module [48].

Previous work by Fiete, Burak and Brookings [64] proposed that the grid system is organized to represent very large ranges in space by exploiting the incommensurability (i.e. lack of common rational factors) of different grid periods. As originally proposed, the grid scales in this scheme were not hierarchically organized (as we now

know they are [164]) but were of similar magnitude, and hence it was particularly important to suggest a scheme where a large spatial range could be represented using grids with small and similar periods. Using all the scales together [64] argued that it is easy to generate ranges of representation that are much larger than necessary for behavior, and Sreenivasan and Fiete argued that the excess capacity could be used for error correction over distances relevant for behavior [162]. However, recent experiments tell us that there is a hierarchy of scales [164] which should make the representation of behaviorally plausible range of 20-100m easily accessible in the alternative hierarchical coding scheme that we have proposed. Nevertheless, we have checked that the optimal grid scheme predicted by our theory, if decoded in the fashion of [64], can represent space over ranges longer than the largest scale (see Supplementary Information) at some excess cost in the number of neurons. It could be that animals sometimes exploit this excess capacity either for error correction or to avoid remapping over a range larger than the period of the largest grid. That said, experiments do tell us that remapping occurs readily over relatively small (meter length) scales at least for dorsal (small scale) place cells and grid cells [73] in tasks that involve spatial cues.

Our hierarchical grid scheme makes distinctive predictions for the effects of selective lesions of grid modules (details in Supplementary Material). We predict that lesioning the modules with small periods will expand place field widths, while lesioning modules with large periods will lead to increased firing at locations outside

the main place field, at scales set by the missing module. Our prediction is supported by a recent study demonstrating effects of lesions including dorsal mEC on place field widths in small environments [87]. Similar effects are predicted for any decoder of a lesioned hierarchical grid system – thus animals with lesions to fine grid modules will show less precision in spatial behavior, while animals with lesions to large grid modules will confound well-separated locations. In contrast, in a non-hierarchical grid scheme with similar but incommensurate periods, lesions of any module lead to the appearance of multiple place fields at many scales for each place cell.

Mathis et al. [126, 125] studied the resolution and representational capacity of grid codes vs. place codes. They found that grid codes have exponentially greater capacity to represent locations than place codes with the same number of neurons. Furthermore, [126] predicted that in one dimension a geometric progression of grids that is self-similar at each scale minimizes the asymptotic error in recovering an animal's location given a fixed number of neurons. Using numerical simulations, they analyzed the dependence of the decoding error on the grid scale factor and found that, in their theory, the optimal scale factor depends on "the number of neurons per module and peak firing rate" and, relatedly, on the "tolerable level of error" during decoding [126]. Our results, which arise from a different formulation of coding precision and resolution, are consistent with these results but additionally allow us to predict structural parameters of the system such as the grid scale factor for spatial coding in different dimensions. These predictions can be directly tested

in experiments.

There is a long history in the study of sensory coding, especially vision, of identifying efficiency principles underlying neural circuits and codes starting with [10]. Our results constitute evidence that such principles might also operate in the organization of cognitive circuits processing non-sensory variables. Furthermore, the existence of an efficiency argument for grid organization of spatial coding suggests that grid systems may be universal amongst the vertebrates, and not just a rodent specialization. In fact, there is evidence that humans [54, 94] and other primates [102] also have grid systems. We expect that our predicted scaling of the grid modules also holds in humans and other primates.

4.4 Supplementary Materials

Optimizing a "base-b" representation of one-dimensional space

Suppose that we want to resolve location with a precision l in a track of length L. In terms of the resolution $R = L/l$, we argued in the Discussion of the main text that a "base-b" hierarchical neural coding scheme will roughly require $N = b \log_b R$ neurons. To derive the optimal base (i.e. the base that minimizes the number of the neurons), we evaluate the extremum $\partial N/\partial b = 0$:

$$\partial N/\partial b = \frac{\partial(b \log_b R)}{\partial b} = \frac{\partial(\frac{b \ln R}{ln b})}{\partial b} = \ln R \, \frac{\ln b - 1}{(\ln b)^2} \tag{4.1}$$

Setting $\partial N/\partial b = 0$ gives $\ln b - 1 = 0$. Therefore the number of neurons is extremized when $b = e$. It is easy to check that this is a minimum. Of course, the base of a number system is usually take to be an integer, so the argument should be taken as motivating the more detailed treatment of neural representations of space in the main text. Neurons, are of course not constrained to organize the periodicity of their tuning curves in integer ratios.

Optimizing the grid system: winner-take-all decoder

Deriving the optimal grid

We saw in the main text that, for a winner-take-all decoder, the problem of deriving the optimal ratios of adjacent grid scales in one dimension is equivalent to minimizing the sum of a set of numbers ($N = d\sum_{i=1}^{m} r_i$) while fixing the product ($R_1 = \prod_{i=1}^{m} r_i$) to take the value R. Mathematically, it is equivalent to minimize N while fixing $\ln R_1$. When N is large we can treat it as a continuous variable and use the method of Lagrange multipliers as follows. First,we construct the auxiliary function $H(r_1 \cdots r_N, \beta) = N - \beta\,(\ln R_1 - \ln R)$ and then extremize H with respect to each r_i and β. Extremizing with respect to r_i gives

$$\frac{\partial H}{\partial r_i} = d - \frac{\beta}{r_i} = 0 \qquad \Longrightarrow \qquad r_i = \frac{\beta}{d} \equiv r\,. \tag{4.2}$$

Next, extremizing with respect to β to implement the constraint on the resolution gives

$$\frac{\partial H}{\partial \beta} = \ln R_1 - \ln R = m \ln r - \ln R = 0 \qquad \Longrightarrow \qquad r = R^{1/m} \tag{4.3}$$

Having thus implemented the constraint that $\ln R_1 = \ln R$, it follows that $H = N = d\,m\,R^{1/m}$. Alternatively, solving for m in terms of r, we can write $H = d\ r\,(\ln R)\,/\,\ln r) = d\,r \log_r R$. It remains to minimize the number of cells N with

respect to r,

$$\frac{\partial H}{\partial r} = d \ln R \left[\frac{1}{\ln r} - \left(\frac{1}{\ln r} \right)^2 \right] = 0 \qquad \Longrightarrow \qquad \ln r = 1 \tag{4.4}$$

This is in turn implies our result

$$r = e \tag{4.5}$$

for the optimal ratio between adjacent scales in a hierarchical, grid coding scheme for position in one dimension, using a winner-take-all decoder.

In this argument we employed the sleight of hand that N and m can be treated as continuous variables, which is approximately valid when N is large. This condition obtains if the required resolution R is large. A more careful argument is given below that preserves the integer character of N and m.

Integer N and m: Above we used Lagrange multipliers to enforce the constraint on resolution and to bound the scale ratios to avoid ambiguity while minimizing the number of neurons required by a Winner-Take-All decoding model of grid systems. Here we will carry out this minimization while recognizing that the number of neurons is an integer. First, consider the arithmetic mean-geometric mean inequality which states that, for a set of non-negative real numbers, $x_1, x_2, ..., x_m$, the following holds:

$$(x_1 + x_2 + ... + x_m)/m \geq (x_1 x_2 ... x_m)^{1/m}\,, \tag{4.6}$$

with equality if and only if all the x_i's are equal. Applying this inequality, it is easy to see that to minimize $\sum_{i=1}^{m} r_i$, all of the r_i should be equal. We denote this common value as r, and we can write $r = R^{1/m}$.

Therefore, we have

$$N = d \sum_{i=1}^{m} r = m\, d\, R^{1/m} \tag{4.7}$$

Suppose $R = e^{z+\epsilon}$, where z is an integer, and $\epsilon \in [0, 1)$. By taking the first derivative of N with respect to m, and setting it to zero, we find that N is minimized when $m = z + \epsilon$. However, since m is an integer the minimum will be achieved either at $m = z$ or $m = z + 1$. (Here we used the fact $mR^{1/m}$ is monotonically increasing between 0 and $z + \epsilon$ and is monotonically decreasing between $z + \epsilon$ and ∞.) Thus, minimizing N requires either

$$r = (e^{z+\epsilon})^{\frac{1}{z}} = e^{\frac{z+\epsilon}{z}} \qquad \text{or} \qquad r = (e^{z+\epsilon})^{\frac{1}{z+1}} = e^{\frac{z+\epsilon}{z+1}}\,. \tag{4.8}$$

In either case, when z is large (and therefore R, N and m are large), $r \to e$. This shows that when the resolution R is sufficiently large, the total number of neurons N is minimized when $r_i \approx e$ for all i.

Optimal winner-take-all grids: general formulation

As described in the main text, we wish to choose the grid system parameters $\{\lambda_i, l_i\}$, $1 \leq i \leq m$, as well as the number of scales m, to minimize neuron number:

$$N = d \sum_{i=1}^{m} \frac{\lambda_i}{l_i}, \tag{4.9}$$

where d is the fixed coverage factor in each module, while constraining the positional accuracy of the grid system and the range of representation. We can take the positional accuracy to be proportional to the grid field width of the smallest module. This gives

$$c_1 \, l_m = A \tag{4.10}$$

To give a sufficiently large range of representation in our hierarchical scheme we will require that

$$\lambda_1 \geq L \tag{4.11}$$

Following the main text, to eliminate ambiguity at each scale we need that

$$\lambda_{i+1} \geq c_2 \, l_i \tag{4.12}$$

where c_2 depends on the tuning curve shape and coverage factor (written as $f(d)$ in the main text).

We will first fix m and solve for the remaining parameters, then optimize over

m in a subsequent step. Optimization problems subject to inequality constraints may be solved by the method of Karush-Kuhn-Tucker (KKT) conditions [109]. We first form the Lagrange function,

$$\mathcal{L} = d \sum_i \frac{\lambda_i}{l_i} + \alpha(c_1\, l_m - A) - \beta_0(\lambda_1 - L) - \sum_{i=1}^{K-1} \beta_i(\lambda_{i+1} - c_2\, l_i). \tag{4.13}$$

The KKT conditions include that the gradient of $\mathcal{L}$ with respect to $\{\lambda_i, l_i\}$ vanish,

$$\frac{\partial \mathcal{L}}{\partial l_m} = c_1\alpha - d\frac{\lambda_m}{l_m^2} = 0 \tag{4.14}$$

$$\frac{\partial \mathcal{L}}{\partial l_i} = c_2\beta_i - d\frac{\lambda_i}{l_i^2} = 0 \quad i < m \tag{4.15}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \frac{d}{l_i} - \beta_{i-1} = 0, \tag{4.16}$$

together with the "complementary slackness" conditions,

$$\beta_0(\lambda_1 - L) = 0 \tag{4.17}$$

$$\beta_i(\lambda_{i+1} - c_2\, l_i) = 0. \tag{4.18}$$

.

From Eqns. (4.15 – 4.16), we obtain:

$$\beta_i = \frac{d}{c}\frac{\lambda_i}{l_i^2} = \frac{d}{l_{i+1}}. \tag{4.19}$$

It follows that $\beta_i \neq 0$, and so the complementary slackness conditions give:

$$\lambda_1 = L \tag{4.20}$$

$$\lambda_i = c_2\, l_{i-1}. \tag{4.21}$$

Substituting this result into Eqn. 4.19 yields,

$$r_i \equiv \frac{l_{i-1}}{l_i} = \frac{l_i}{l_{i+1}} = r_{i+1}, \tag{4.22}$$

i.e., the scale factor r is the same for all modules. Once we obtain a value for r, Eqns. (4.20 – 4.22) yield values for all λ_i and l_i. Since the resolution constraint may now be rewritten,

$$A = c_1 r^{-m} L, \tag{4.23}$$

we have $m = \ln(c_1\, L/A)/\ln r$. Therefore, r determines m and so minimizing N over m is equivalent to minimizing over r. Expressing N entirely in terms of r gives,

$$N = d\, c_2 \ln(c_1 L/A)\frac{\ln r}{r}. \tag{4.24}$$

Optimizing with respect to r gives the result $r = e$, independent of d, c_1, c_2, L, and R.

Optimizing the grid system: probabilistic decoder

Consider a probabilistic decoder of the grid system that pools all the information available in the population of neurons in each module by forming the posterior distribution over position given the neural activity. In this general setting, we assume that the firing of different grid cells is weakly correlated, that noise is homogeneous, and that the tuning curves in each module i provide dense, uniform, coverage of the interval λ_i. With these assumptions, we will first consider the one-dimensional case, and then analyze the two-dimensional case by analogy.

One-dimensional grids: With the above assumptions, the likelihood of the animal's position, given the activity of grid cells in module i, $P(x \mid i)$, can be approximated as a series of Gaussian bumps of standard deviation σ_i spaced at the period λ_i [48]. As defined in the main text, the number of cells (n_i) in the ith module, is expressed in terms of the period (λ_i), the grid field width (l_i) and a "coverage factor" d representing the cell density as $n_i = d\lambda_i/l_i$. The coverage factor d will control the relation between the grid field width l_i and the standard deviation σ_i of the local peaks in the likelihood function of location. If d is larger, σ_i will be narrower since we can accumulate evidence from a denser population of neurons. The ratio $\frac{l_i}{\sigma_i}$ in general will be a monotonic function of the coverage factor d, which we will write as $\frac{l_i}{\sigma_i} = g(d)$. In the special case where the grid cells have independent noise $g(d) \propto \sqrt{d}$, so that $\sigma_i/l_i \propto 1/\sqrt{d}$ – i.e. the precision increases as the inverse

square root of the cell density, as expected because the relevant parameter is the number of cells within one grid field rather than the total number of cells. Note that this does *not* imply an inverse square root relation between the *number* of cells n_i and σ_i, because n_i is also proportional to the period λ_i, and in our formulation the density d is fixed while λ_i can be varied. Note also that if the neurons have correlated noise, $g(d)$ may scale substantially slower than $\sqrt{d}$ [24, 188, 161]. Putting all of these statements together, we have, in general, $n_i = \frac{d}{g(d)} \frac{\lambda_i}{\sigma_i}$. Assuming that the coverage factor d is the same across modules, we can simplify the notation and write $n_i = c\frac{\lambda_i}{\sigma_i}$, where $c = d/g(d)$ is a constant. (Note again that for independent noise $\sigma_i \propto 1/\sqrt{d}$ as expected – see above – and this does *not* imply a similar relationship to the number of cells n_i as one might have naively assumed.) In sum, we can write the total number of cells in a grid system with m modules as $N = \sum_{i=i}^{m} n_i = c \sum_{i=1}^{m} \frac{\lambda_i}{\sigma_i}$.

The likelihood of position derived from each module can be combined to give an overall probability distribution over location. Let $Q_i(x)$ be the likelihood obtained by combining modules 1 (the largest period) through i. Assuming that the different modules have independent noise, we can compute $Q_i(x)$ from the module likelihoods as $Q_i(x) \propto \prod_{j=1}^{n} P(x|j)$. We will take the prior probability over locations be uniform here so that this combined likelihood is equivalent to the Bayesian posterior distribution over location. The likelihoods from different scales have different periodicities, so multiplying them against each other will tend to suppress all peaks

except the central one, which is aligned across scales. We may thus approximate $Q_i(x)$ by single Gaussians whose standard deviations we will denote as δ_i. (The validity of this approximation is taken up in further detail below.)

Since $Q_i(x) \propto Q_{i-1}(x)P(x|i)$, δ_i is determined by δ_{i-1}, λ_i and σ_i. These all have dimensions of length. Dimensional analysis [140] therefore says that, without loss of generality, the ratio δ_i/δ_{i-1} can be written as a dimensionless function of any two cross-ratios of these parameters. We can use this freedom to write $\delta_i = \delta_{i-1}/\rho(\frac{\lambda_i}{\sigma_i}, \frac{\sigma_i}{\delta_{i-1}})$. The standard error in decoding the animal's position after combining information from all the grid modules will be proportional to δ_m, the standard deviation of Q_m. We can iterate our expression for δ_i in terms of δ_{i-1} to write $\delta_m = (\prod_{i=1}^m \rho_i)\,\delta_0$ where δ_0 is the uncertainty in location without using any grid responses at all. (We are abbreviating $\rho_i = \rho(\lambda_i/\sigma_i, \sigma_i/\delta_{i-1})$. In the present probabilistic context, we can view δ_0 as the standard deviation of the *a priori* distribution over position before the grid system is consulted, but it will turn out that the precise value or meaning of δ_0 is unimportant. We assume a behavioral requirement that fixes δ_m and thus the resolution of the grid, and that δ_0 is likewise fixed by the behavioral range. Thus, there is a constraint on the product $\prod_i \rho_i$.

Putting everything together, we wish to minimize $N = c\sum_{i=1}^m \frac{\lambda_i}{\sigma_i}$ subject to the constraint that $R = \prod_{i=1}^m \rho_i$ where ρ_i is a function of λ_i/σ_i and σ_i/δ_{i-1}. Given the formula for ρ_i derived in the next section, this can be carried out numerically. To understand the optimum it is helpful to observe that the problem has a symmetry

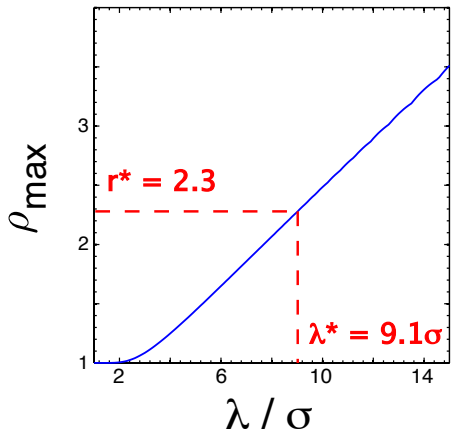

Figure 4.5: Optimizing the one dimensional grid system. $\rho_{max} \equiv \max_{\sigma/\delta} \rho(\frac{\lambda}{\sigma}, \frac{\sigma}{\delta})$ is the scale factor after optimizing N over σ/δ. The values r^* and λ^* are the values chosen by the complete optimization procedure.

under permutations of i. So we can guess that in the optimum all the λ_i/σ_i, σ_i/δ_{i-1} and ρ_i will be equal to a fixed λ/σ, σ/δ and ρ. We can look for a solution with this symmetry and then check that it is an optimum. First, using the symmetry, we write $N = c\,m(\lambda/\sigma)$ and $R = \rho^m$. It follows that $N = c(1/\ln\rho)(\lambda/\sigma)$ and we want to minimize it with respect to λ/σ and σ/δ. Now, $\rho(\lambda/\sigma, \sigma/\delta)$ is a complicated function of its arguments (4.30) which has a maximum value as a function of σ/δ for any fixed λ/σ. To minimize N at fixed λ/σ we should maximize ρ (Fig. 4.5). Given this ρ_{max}, we can minimize $N = c\,(\lambda/\sigma)/\ln\rho_{max}(\lambda/\sigma)$ with respect to λ/σ, and then plug back in to find the optimal ρ. It turns out to be $\rho^*_{max} = 2.3$

In fact, ρ is equal to the scale factor of the grid: $\rho_i = r_i = \lambda_i/\lambda_{i+1}$. To see this, we have to express ρ_i in terms of the parameters λ_i/σ_i and σ_i/δ_{i-1}: $\rho_i = \frac{\delta_{i-1}}{\delta_i} = \frac{\delta_{i-1}}{\sigma_i}\frac{\sigma_i}{\lambda_i}\frac{\lambda_i}{\lambda_{i+1}}\frac{\lambda_{i+1}}{\sigma_{i+1}}\frac{\sigma_{i+1}}{\delta_i}$. Since the factors σ_i/δ_{i-1} and λ_i/σ_i are independent of i, they cancel in the product and we are left with $\rho_i = \lambda_i/\lambda_{i+1}$.

Thus the probabilistic decoder predicts an optimal scale factor $r^* = 2.3$ in one dimension. This is similar to, but somewhat different than, the winner-take-all result $r^* = e = 2.7$. At a technical level, the difference arises because the function $\rho_{max}(\lambda/\sigma)$ is effectively $\rho_{max} \propto \frac{\lambda}{\sigma}$ in the winner-take-all analysis, but in the the probabilistic case it is more nearly a linear function with a positive offset $\rho \approx \alpha^{-1}(\frac{\lambda}{\sigma} + \beta)$. Conceptually, the optimal probabilistic scale factor is smaller in order to suppress side lobes that can arise in the combined likelihood across modules (Fig. 2 of the main text). Such side lobes were absent in the winner-take-all analysis. The optimization also predicts $\lambda^* = 9.1\,\sigma$. This relation between the period and standard deviation at each scale could be converted into a relation between grid period and grid field width given specific measurements of tuning curves, noise levels, and cell density in each module. For example, if neurons within a module have independent noise, then general population coding considerations [48] show that $\sigma = g\,d^{-1/2}l$ where l is a measure of grid field width, d is the density of neurons in a module, and g is a dimensionless number that depends on noise (given the integration time) and tuning curve shape.

Two dimensional grids: A similar probabilistic analysis can be carried out for two dimensional grid fields. The posteriors $P(x \,|\, i)$ become two-dimensional sums-of-Gaussians, with the centers of the Gaussians laid out on the vertices of the grid. $Q_i(x)$ is then similarly approximated by a two-dimensional Gaussian. Generalizing from the one-dimensional case, the number of cells in module i is

given by $n_i = d(\lambda_i/l_i)^2$, where d is density of grid fields. As in one dimension, increasing the density d will decrease the standard deviation σ_i of the local bumps in the posterior $P(x \mid i)$ – i.e., $l_i/\sigma_i = g(d)$ where g is an increasing function of d. In the special case where the neurons have independent noise, $g(d) \propto \sqrt{d}$ so that the precision measured by the standard deviation σ_i decreases as the inverse square root of d. Putting all of these statements together, we have, in general, $n_i = \frac{d}{g(d)^2}\left(\frac{\lambda_i}{\sigma_i}\right)^2$. In the special case where noise is independent so that $g(d) \propto \sqrt{d}$ the density d cancels out in this expression, and in this case, or when the density d is the same across modules, we can write $n_i = c\left(\frac{\lambda_i}{\sigma_i}\right)^2$ where c is just a constant. Redoing the optimization analysis from the one dimensional case, the form of the function ρ changes (section 4.3), but the logic of the above derivation is otherwise unaltered. In the optimal grid we find that $\lambda^* \approx 5.3\sigma$ (or equivalently $\sigma \approx 0.19\lambda^*$).

Calculating $\rho(\frac{\lambda}{\sigma}, \frac{\sigma}{\delta})$

Above, we argued that the function $\rho(\frac{\lambda}{\sigma}, \frac{\sigma}{\delta})$ can be computed by approximating the posterior distribution of the animal's position given the activity in module i, $P(x \mid i)$, as a periodic sum-of-Gaussians:

$$P(x \mid i) = \frac{1}{2K+1} \sum_{n=-K}^{K} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(x-n\lambda_i)^2} \tag{4.25}$$

where K is assumed large. We further approximate the posterior given the activity of *all* modules coarser than λ_i by a Gaussian with standard deviation δ_{i-1}:

$$Q_{i-1}(x) = \frac{1}{\sqrt{2\pi\delta_{i-1}^2}} e^{-x^2/2\delta_{i-1}^2} \tag{4.26}$$

(We are assuming here that the animal is really located at $x = 0$ and that the distributions $P(x|i)$ for each i have one peak at this location.) Assuming noise independence across scales, it then follows that $Q_i(x) = \frac{P(x\,|\,i)Q_{i-1}(x)}{\int \mathrm{d}x P(x\,|\,i)Q_{i-1}(x)}$. Then $\rho(\lambda_i/\sigma_i, \sigma_i/\delta_{i-1})$ is given by δ_{i-1}/δ_i, where δ_i is the standard deviation of Q_i. We therefore must calculate $Q_i(x)$ and its variance in order to obtain ρ. After some algebraic manipulation, we find,

$$Q_i(x) = \sum_{n=-K}^{K} \pi_n \frac{1}{\sqrt{2\pi\Sigma^2}} e^{-(x-\mu_n)^2/2\Sigma^2}, \tag{4.27}$$

where $\Sigma^2 = \left(\sigma_i^{-2} + \delta_{i-1}^{-2}\right)^{-1}$, $\mu_n = \left(\frac{\Sigma}{\sigma_i}\right)^2 \lambda_i\, n$, and

$$\pi_n = \frac{1}{Z} e^{-n^2\lambda_i^2/2(\sigma_i^2+\delta_{i-1}^2)}. \tag{4.28}$$

Z is a normalization factor enforcing $\sum_n \pi_n = 1$. Q_i is thus a mixture-of-Gaussians, seemingly contradicting our approximation that all the Q are Gaussian. However, if the secondary peaks of $P(x\,|\,i)$ are well into the tails of $Q_{i-1}(x)$, then they will be suppressed (quantitatively, if $\lambda_i^2 \gg \sigma_i^2 + \delta_{i-1}^2$, then $\pi_n \ll \pi_0$ for $|n| \geq 1$), so that

our assumed Gaussian form for Q holds to a good approximation. In particular, at the values of λ, σ, and δ selected by the optimization procedure described above, $\pi_1 = 1.3 \cdot 10^{-3}\pi_0$. So our approximation is self-consistent.

Next, we find the variance δ_i^2:

$$\begin{aligned}
\delta_i^2 &= \langle x^2 \rangle_{Q_i} \\
&= \sum_n \pi_n (\Sigma^2 + \mu_n^2) \\
&= \Sigma^2 \left(1 + \left(\frac{\Sigma}{\sigma_i} \right)^2 \left(\frac{\lambda_i}{\sigma_i} \right)^2 \sum_n n^2 \pi_n \right) \\
&= \delta_{i-1}^2 \left(1 + \frac{\delta_{i-1}^2}{\sigma_i^2} \right)^{-1} \left(1 + \left(\frac{\Sigma}{\sigma_i} \right)^2 \left(\frac{\lambda_i}{\sigma_i} \right)^2 \sum_n n^2 \pi_n \right).
\end{aligned} \tag{4.29}$$

We can finally read off $\rho(\frac{\lambda_i}{\sigma_i}, \frac{\sigma_i}{\delta_{i-1}})$ as the ratio δ_{i-1}/δ_i:

$$\rho(\frac{\lambda_i}{\sigma_i}, \frac{\sigma_i}{\delta_{i-1}}) = \left(1 + \frac{\delta_{i-1}^2}{\sigma_i^2} \right)^{1/2} \left(1 + \left(1 + \frac{\sigma_i^2}{\delta_{i-1}^2} \right)^{-1} \left(\frac{\lambda_i}{\sigma_i} \right)^2 \sum_n n^2 \pi_n \right)^{-1/2}. \tag{4.30}$$

For the calculations reported in the text, we took $K = 500$.

We explained above that we should maximize ρ over $\frac{\sigma}{\delta}$, while sholding $\frac{\lambda}{\sigma}$ fixed. The first factor in Eq. 4.30 increases monotonically with decreasing $\frac{\sigma}{\delta}$; however, $\sum_n n^2 \pi_n$ also increases and this has the effect of reducing ρ. The optimal $\frac{\sigma}{\delta}$ is thus controlled by a tradeoff between these factors. The first factor is related to the increasing precision given by narrowing the central peak of $P(x \,|\, i)$, while the second factor describes the ambiguity from multiple peaks.

Generalization to two dimensional grids: The derivation can be repeated in the two-dimensional case. We take $P(x \mid i)$ to be a sum-of-Gaussians with peaks centered on the vertices of a regular lattice generated by the vectors $(\lambda_i \hat{u}, \lambda_i \vec{v})$. We also define $\delta_i^2 \equiv \frac{1}{2}\langle |x|^2 \rangle_{Q_i}$. The factor of 1/2 ensures that the variance so defined is measured as an average over the two dimensions of space. For analytical simplicity we also assumed that the grid orientations were aligned across modules. The derivation is otherwise parallel to the above, and the result is,

$$\rho_2(\frac{\lambda_i}{\sigma_i}, \frac{\sigma_i}{\delta_{i-1}}) = \left(1 + \frac{\delta_{i-1}^2}{\sigma_i^2}\right)^{1/2} \left(2 + \left(1 + \frac{\sigma_i^2}{\delta_{i-1}^2}\right)^{-1} \left(\frac{\lambda_i}{\sigma_i}\right)^2 \sum_{n,m} |n\hat{u} + m\vec{v}|^2 \ \pi_{n,m}\right)^{-1/2}, \tag{4.31}$$

where $\pi_{n,m} = \frac{1}{Z} e^{-|n\hat{u}+m\vec{v}|^2 \lambda_i^2 / 2(\sigma_i^2 + \delta_{i-1}^2)}$.

Reanalysis of grid data from previous studies

We reanalyzed the data from Barry et. al [11] and Stensola et al. [164] in order to get the mean and the variance of the ratio of adjacent grid scales. For Barry et al. [11], we first read the raw data from Figure 3b of the main text using the software GraphClick, which allows retrieval of the original (x,y)-coordinates from the image. This gave the scales of grid cells recorded from 6 different rats. For each animal, we grouped the grids that had similar periodicities (i.e. differed by less than 20%) and calculated the mean periodicity for each group. We defined this mean periodicity as the scale of each group. For 4 out of 6 rats, there were 2 scales

in the data. For 1 out 6 rats, there were 3 grid scales. For the remaining rat, only 1 scale was obtained as only 1 cell was recorded from that rat. We excluded this rat from further analysis. We then calculated the ratio between adjacent grid scales, resulting in 6 ratios from 5 rats. The mean and variance of the ratio were 1.64 and 0.09, respectively ($n = 6$).

For Stensola et. al[164], we first read in the data using GraphClick from Figure 5d of the main text. This gave the scale ratios between different grids for 16 different rats. We then pooled all the ratios together and calculated the mean and variance. The mean and variance of the ratio were 1.42 and 0.17, respectively ($n = 24$).

Giocomo et. al[81] reported the ratios between the grid period and the *radius* of grid field (measured as the radius of the circle around the center field of the autocorrelation map of the grid cells) to be 3.26 ± 0.07 and 3.32 ± 0.06 for Wild-type and HCN KO mice, respectively. We halved these measurements to the ratios between grid period and the *diameter* of the grid field to facilitate the comparison to our theoretical predictions. The results are plotted in a bar graph (Fig. 4B in the main text).

Finally, in Figure 4C, we replotted Fig. 1c from [86] by reading in the data using GraphClick and then translating that information back into a plot.

Range of location coding in a grid system

The main text describes hierarchical grid coding schemes where the larger periods resolve ambiguity and smaller periods give precision in location coding. We took the largest grid period to be comparable to the behavioral range. In fact, if the periods λ_i of the different modules are incommensurate with each other (i.e. they do not share common integer factors), it should be possible to resolve location over ranges larger than the largest grid period [64, 162]. The grid schemes that we predict share this virtue since they predict scale ratios that are not simple rational numbers. However the precise maximum range will also depend on the widths of the grid fields l_i relative to the period and on the number of grid cells n_i in each module. In the probabilistic decoding scheme described in the main text, these parameters determine the standard deviation σ_i of the periodic peaks in the likelihood of position given the activity in module i. The full range of unambiguous location representation depends on the ratios λ_i/σ_i. Increasing this ratio will tend to increase the range of unambiguous representation, but at the cost of increasing the number of cells in each module.

To illustrate, consider a one-dimensional grid system with 4 modules with a ratio of 2.7 between adjacent scales (this is close to the optimal ratio predicted by our analysis). Suppose the animal's true location is at 0. We can calculate the overall probability of the animal's location by multiplying together the likelihood functions resulting from activity in each individual module (see main text for details). We

will examine the extent to which location can be decoded unambiguously over a range $(-3\lambda_{max}, 3\lambda_{max})$ where $\lambda_{\max}$ is the larges period. When λ_i/σ_i is close to the value of 9.1 predicted by the probabilistic analysis in Sec. 4.3, the overall likelihood shows substantial ambiguity over this range because of secondary peaks in the likelihood distribution (Fig. 4.6A). As λ_i/σ_i increases (requiring more neurons in each module), these secondary peaks decrease in amplitude. In Fig. 4.6B, we show that when $\lambda_i/\sigma_i = 30$, the 4-module grid system can represent location at least within the range $(-3\lambda_{max}, 3\lambda_{max})$.

If there is a biological limitation to the largest period possible in a grid system, and if the organism must represent very large ranges without grid remapping, it may prove beneficial to add neurons to expand range. Analyzing this tradeoff requires knowledge of the range, biophysical limits on grid periods, and the degree of ambiguity (the maximum heights of secondary peaks in the probability of position) that can be behaviorally tolerated. This information is not currently available for any species, and so we do not attempt the analysis.

Predictions for the effects of lesions and for place cell activity

In the grid coding scheme that we propose there is a hierarchy of grid periods governed by a geometric progression. The alternative schemes of [64, 162] are

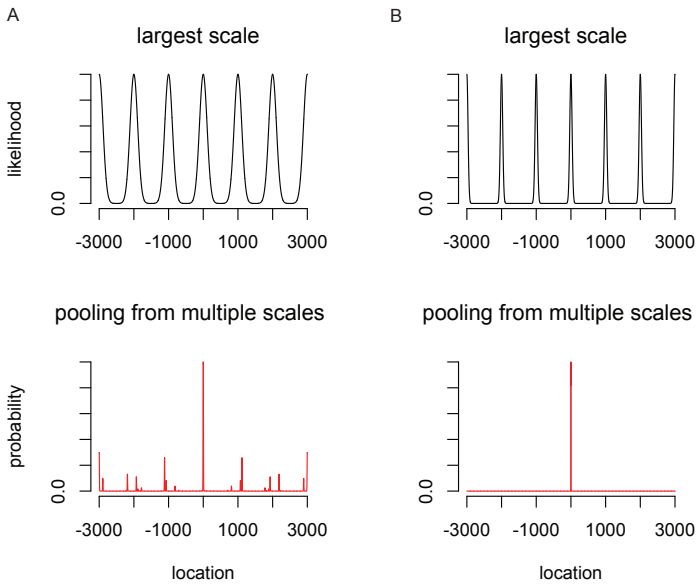

Figure 4.6: Encoding range can exceeds the period of the largest grid module. Assume that the animal is located is at 0. (**A**) Top: The likelihood resulting from the largest grid module, where the standard deviation of the Gaussian peaks is $\frac{1}{9.1}$ of the grid period ($\lambda_{max} = 1000$). Bottom: The inferred distribution over location after pooling over 4 grid modules related by a scale factor of 2.7. As shown, this 4-module grid system shows ambiguities in location coding outside the range $[\lambda_{max}, \lambda_{max}]$. (**B**) Top: The likelihood resulting from the largest grid module, where the standard deviation of the Gaussian peaks is $\frac{1}{30}$ of the grid period ($\lambda_{max} = 1000$). Bottom: The inferred distribution over location after pooling over 4 grid modules related by a scale factor of 2.7. As shown, this 4-module grid system provides a good representation over a range of at least $[-3000, 3000] = [-3\lambda_{max}, 3\lambda_{max}]$.

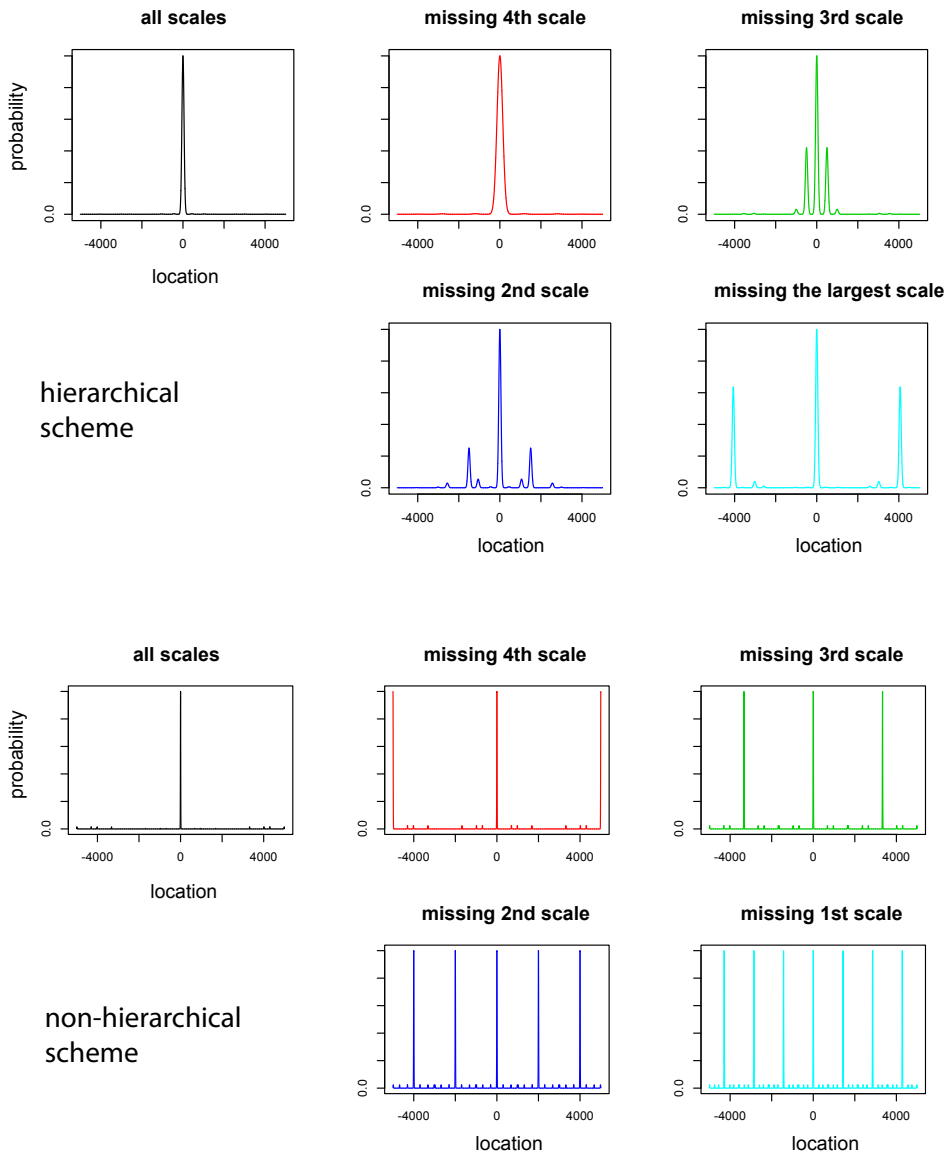

Figure 4.7: The effect of lesioning grid modules on the distribution over location for hierarchical vs. non-hierarchical grid schemes. For the hierarchical scheme, we assume that 4 one-dimensional grid modules are related by a scale factor r ($r = 2.7$), i.e. $\frac{\lambda_i}{\lambda_{i+1}} = 2.7$, $i = 1, 2, 3$, and the ratio $\frac{\lambda_i}{\sigma_i} = 9.1$, i=1,2,3,4. We assume that the animal is at $x = 0$ and construct the probability distribution over location given the activity in each grid module as described in Sec. 4.3. For the non-hierarchical scheme, we again assume 4 grid modules and set the periods of the 4 modules to be 1/105 (4^{th}), 1/70 (3^{rd}), 1/42 (2^{nd}),1/30 (1^{st}) of the whole range respectively. We set the width of the composite likelihood after combining all 4 modules to be 1/210 of the range $[-5000, 5000]$.

designed to produce a large range of representation from grids with *similar* periods. These two alternatives make very different predictions for the effects of lesions in the entorhinal cortex on location coding. In a hierarchical scheme, losing a grid module produces location ambiguities that increase in size with the the period of the missing module. In the alternative scheme of [64, 162] lesions of a module produce periodic ambiguities that are sporadically tied to the missing period. An illustrative example is shown in Fig. 4.7.

The grid cell representation of space in the entorhinal cortex is thought to be transformed in the hippocampus into the place cell representation. Simple models of this transformation assume that grid cells are pooled in the hippocampus and that some form of synaptic plasticity selects inputs with the same spatial phase [160]. In the context of such a model, our grid scheme makes specific predictions for the effects of module lesions on place fields.

We use a firing rate model for both place cells and grid cells. The 1-d grid cell firing rate is modeled as a periodic sum of truncated Gaussians (a full Gaussian mixture model gives similar results but the truncated model is easier to handle numerically). We will consider four grid modules with module periods λ_i, Gaussian standard deviations σ_i of the bump of the grid cell tuning curve, and ratios $\lambda_i/\sigma_i = 9.1$. The grid periods follow a scaling $\lambda_i/\lambda_{i+1} = 2.7$, and we examine place coding over the range set by the biggest period λ_1.

The place cell response is modeled via linear pooling of grid cells with the same

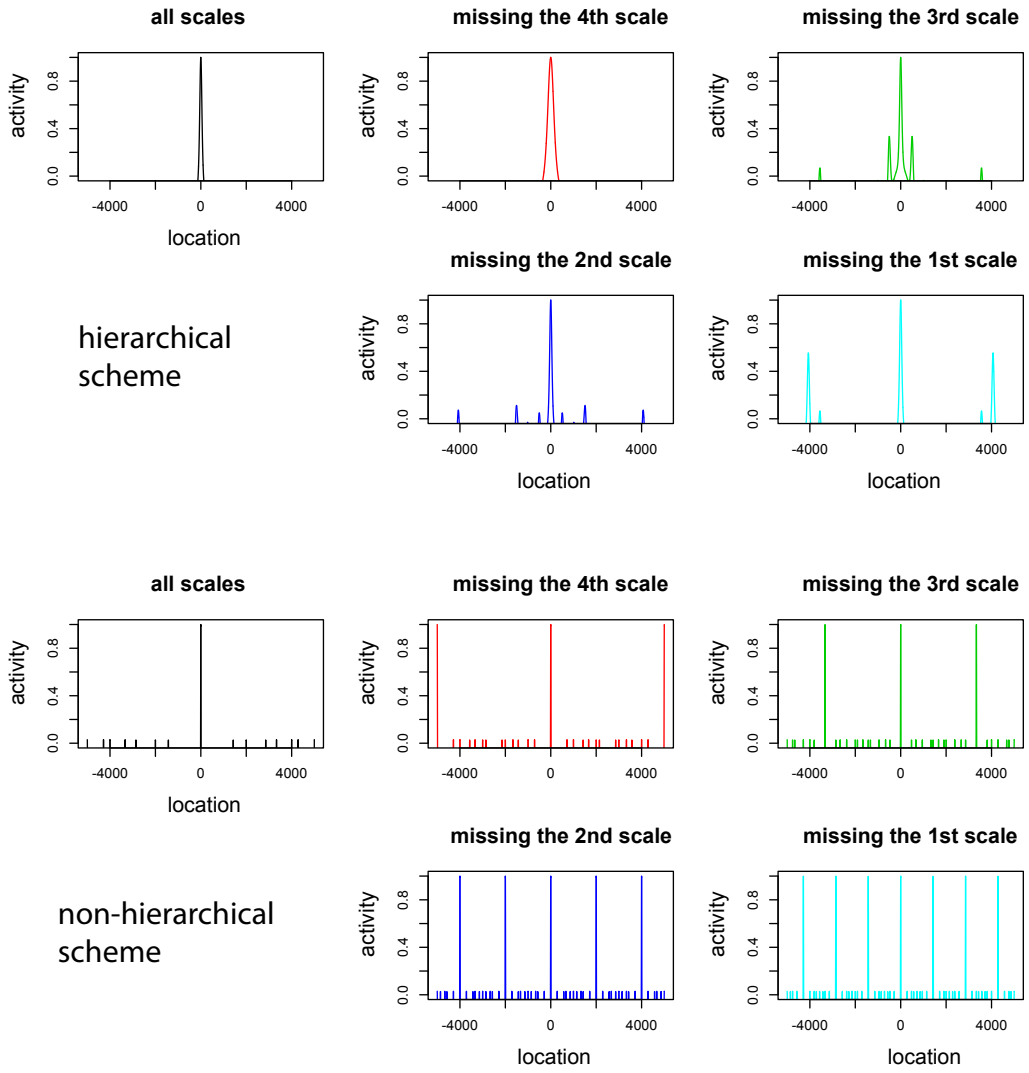

Figure 4.8: The effect of lesioning individual grid modules on place cell activity in a simple grid-place transformation model. Lesioning different modules lead to qualitatively different effects on the place cell response in the hierarchical coding scheme we proposed, as compared to a non-hierarchical scheme. See Sec. 4.3 for details.

phase followed by a threshold and an exponential nonlinearity:

$$f(x) \propto exp(\sum_{1}^{4} g_i(x)) - c * m.$$

Here $g_i(x)$ is the grid cell firing rate, $c = 0.3$ sets the threshold and $m = max\ \{exp(\sum_1^4 g_i(x))\}$ is the maximum activation. This is a simplified description of the essential features of many models of the grid-place transformation (see, e.g., [49, 160] and the review [82]). To model the effect of lesioning grid module i, we set the $g_i(x) = 0$. The results are shown in the Fig. 4.8. Qualitatively, lesioning the smallest grid module increases the place cell width, while lesioning the largest grid module leads to increased firing in locations outside the main place fields. In general, lesioning different grid modules along the hierarchy leads to different effects on the place field. This is a testable prediction in future experiments. Note that lesions of dorsal-ventral bands are not a direct test – multiple grid modules co-exist in each location along the dorsal-ventral axis [164].

For comparison purposes, we also simulated a non-hierarchical model where grid periods are similar but incommensurate. In this model, the place cell response is

$$\tilde{f}(x) \propto exp(\sum_{1}^{4} \tilde{g}_i(x)) - \tilde{c} * \tilde{m},$$

where $c = \tilde{0}.35$ is a threshold, $\tilde{m} = max\ \{exp(\sum_1^4 \tilde{g}_i(x))\}$, and $\tilde{g}_i(x)$ is the grid cell firing rate again modeling as a sum of truncated Gaussians. In each module we took

the standard deviation of the Gaussians to be 1/210 of the whole range. The periods of the grids in the four modules were 1/105 (4^{th}), 1/70 (3^{rd}), 1/42 (2^{nd}),1/30 (1^{st}) of the whole range respectively. Again, to model the effect of lesioning grid module i, we set the $\tilde{g}_i(x) = 0$. In this grid scheme, lesioning any grid module leads to qualitatively similar effects on the place cell activity, as they all lead to the emergence of several place fields (Fig. 4.8). This is in contrast with the hierarchical scheme, in which lesioning the largest scale leads to an expansion of place fields rather than an increase in the number of fields.

Chapter 5

Mutual information, Fisher information, Efficient coding

5.1 Introduction

The Efficient coding hypothesis is an important proposal of how neural systems may represent (sensory) information [10, 4, 121]. Common formulations of Efficient coding are based on the assumption that a neural system is adapted to the statistical structure of the environment such that the mutual information [153] between the stimulus variable and its neural representation (*e.g.*, as reflected in the firing activity of a neural population) is maximized subject to certain resource constraints. However, the test of this prominent hypothesis is impeded by the fact that mutual information is analytically tractable only for simple coding problems [111, 3].

One way to work around this difficulty is to relate mutual information to another, more tractable information quantity such as Fisher Information [67]. For many neural population coding models, Fisher Information is relatively easy to compute and to interpret with regard to neurophysiological parameters (*e.g.*, neural response gain and dynamic range) as well as psychophysical behavior (*e.g.*, discrimination threshold [152, 151]). In a seminal paper, Brunel and Nadal [27] argued that Fisher information provides a lower bound on mutual information. This result has been widely applied in various studies aimed at testing the Efficient coding hypothesis *e.g.*, [88, 127, 76, 181, 77]. However, some recent theoretical and numerical analyses raise doubts on whether Fisher information indeed represents a lower bound on mutual information [15, 185].

In light of these conflicting results, we revisited the formal link between Fisher and mutual information. We first re-examined the conditions for which the lower bound proposed by [27] holds. We show that the derivation of the bound is based on assumptions that make it automatically tight, thus defying the meaning of a bound. We then formally derive the relation between Fisher and mutual information in a standard input-output model under more general conditions. We discuss the possible interpretation of this relation in terms of both, upper and lower bounds on mutual information. Finally, we demonstrate the implications of our result for understanding neural coding characteristics. In particular, we derive neural and behavioral signatures of Efficient coding.

5.2 Examining the derivation of a lower bound on mutual information

Brunel and Nadal [27] derived a lower bound on the mutual information contained in a neural code, using Fisher information. Neural coding was formulated as a channel coding problem where an input (stimulus) variable θ is encoded in the output (measurement) m of a noisy channel. Denote the mutual information between a stimulus variable θ and the sensory measurement m to be $I[\theta, m]$. Rather than directly computing this quantity, Brunel and Nadal considered the mutual information between θ and $\hat{\theta}$, where $\hat{\theta}$ is the output of an unbiased efficient estimator with mean θ and variance $1/J(\theta)$, and

$$J(\theta) = \int \left(\frac{\partial \ln p(m|\theta)}{\partial \theta} \right)^2 p(m|\theta) dm \tag{5.1}$$

is the Fisher information of the estimate with regard to the input. The mutual information between θ and $\hat{\theta}$ then can be written as

$$I[\theta, \hat{\theta}] = H[\hat{\theta}] - \int d\theta p(\theta) H[\hat{\theta}|\theta]. \tag{5.2}$$

The moment-entropy inequality [39] states that for a continuous random variable with given variance, the Shannon entropy [153] is maximal if and only if the variable

is Gaussian distributed. Thus we can consider

$$H[\hat{\theta}|\theta] \leq H[Z] = \frac{1}{2}\ln\left(\frac{2\pi e}{J(\theta)}\right), \tag{5.3}$$

where Z is a Gaussian random variable with mean θ and variance $1/J(\theta)$, and rewrite the mutual information as the inequality

$$I[\theta,\hat{\theta}] \geq H[\hat{\theta}] - \int d\theta p(\theta)\frac{1}{2}\ln\left(\frac{2\pi e}{J(\theta)}\right) \ . \tag{5.4}$$

Due to the data processing inequality, $I[\theta, m] \geq I[\theta,\hat{\theta}]$. Assuming the asymptotic limit $H[\hat{\theta}] \to H[\theta]$, we finally arrive at

$$I[\theta, m] \geq \underbrace{H[\theta] - \int d\theta p(\theta)\frac{1}{2}\ln\left(\frac{2\pi e}{J(\theta)}\right)}_{I_{\text{Fisher}}}, \tag{5.5}$$

which states that Fisher information provides a lower bound on mutual information.

Limitations of the formulation Although the derivation of the lower bound Eq. (5.5) is correct, it relies on assumptions that strongly compromise its interpretation as a bound. First, there is the assumption that an *unbiased efficient estimator* $\hat{\theta}(m)$ exists. This assumption implies that the noise model must be a member of the exponential family, and θ has to be the natural parameter of the particular exponential family [115, 1]. Second, the lower bound is derived for the asymptotic

limit where any noise model from the exponential family is largely equivalent to a Gaussian model [115]. These two assumptions essentially require the noise to be small and Gaussian, in which case the bound Eq. (5.5) is tight and becomes an identity [33, 145]. Thus, while the identity is with no doubt an interesting and useful result, the notion that Eq. (5.5) represents a lower bound on mutual information, however, seems not particularly meaningful.

5.3 A new look at the link between mutual information and Fisher information

In the following, we revisit the formal link between Fisher and mutual information, in particular with regard to non-Gaussian noise models that are often relevant for the assessment of neural codes.

For analytical convenience, we will consider a standard one-dimensional input-output model [111, 131, 13] between the sensory variable θ and its neural representation m. More specifically, we assume

$$m = f(\theta) + \delta, \tag{5.6}$$

where θ has a continuous prior distribution $p(\theta)$, $f(\theta)$ is an invertible *transfer function* that is bounded, and δ represents arbitrary additive noise with smooth density $q(.)$.

5.3.1 Stam's inequality

We first introduce *Stam's inequality* [163] that which is often applied in information theory yet is little known in the neural coding literature. The inequality plays an important role in the derivation of our main result. We begin by reformulating Fisher information with regard to our input-output model (Eq. (5.6)). With $\tilde{\theta} = f(\theta)$, Fisher information with respect to $\tilde{\theta}$ is given as

$$J(\tilde{\theta}) = \int \left(\frac{\partial \ln p(m|\tilde{\theta})}{\partial \tilde{\theta}} \right)^2 p(m|\tilde{\theta}) dm \ . \tag{5.7}$$

Because we assume additive noise with density $q(.)$, we can write $p(m|\tilde{\theta}) = q(m - \tilde{\theta})$. In this case, $J(\tilde{\theta})$ becomes independent of $\tilde{\theta}$ and thus constant [163], and can be rewritten as

$$J[\delta] := \int \left(\frac{\partial \ln \ q(\delta)}{\partial \delta} \right)^2 q(\delta) d\delta. \tag{5.8}$$

This quantity is referred to as Fisher information of a random variable with respect to a scalar translation parameter [51]. Note that we use a different notation $J[\delta]$ in order to distinguish it from the standard formulation of Fisher information $J(\tilde{\theta})$. Conceptually, $J[\delta]$ summarizes the total local dispersion of a distribution.

The Shannon entropy $H[m|\tilde{\theta}]$ [153] is also independent of $\tilde{\theta}$ and identical to the noise entropy

$$H[\delta] = -\int q(\delta) \ln q(\delta) d\delta. \tag{5.9}$$

Stam's inequality specifies the relation between Fisher information $J[\delta]$ and Shannon entropy $H[\delta]$ as the following:

> *For a given amount of Fisher information, the Shannon entropy of a continuous random variable is minimized if and only if the variable is Gaussian distributed. [163]*

Thus with the notation above (Eq. (5.8 and (5.9)), Stam's inequality implies that

$$H[\delta] \geq \frac{1}{2} \ln \left(\frac{2\pi e}{J[\delta]} \right) \ . \tag{5.10}$$

For Gaussian distributed δ with variance σ^2, Shannon entropy and Fisher information are $H[\delta] = \frac{1}{2} \ln(2\pi e \sigma^2)$ and $J[\delta] = 1/\sigma^2$, respectively. As a remark, Eq. (5.10) is equivalent to isoperimetric inequality for entropies in the information theory literature [51].

5.3.2 Main result

With the above results, we can now express mutual information in terms of Fisher information. Because the transfer function f is invertible, $I[\theta, m] = I[\tilde{\theta}, m]$. Thus we can write mutual information as

$$I[\tilde{\theta}, m] = H[m] - \int d\tilde{\theta} p(\tilde{\theta}) H[m|\tilde{\theta}] \ . \tag{5.11}$$

As we have shown in Section 5.3.1, $H[m|\tilde{\theta}] = H[\delta]$, and thus

$$I[\tilde{\theta}, m] = H[m] - \int d\tilde{\theta} p(\tilde{\theta}) H[\delta] \ . \tag{5.12}$$

Defining D_0 to be the entropy difference between the noise δ and a Gaussian with the same amount of Fisher information $J[\delta]$, *i.e.*,

$$D_0 = H[\delta] - \frac{1}{2} \ln \left(\frac{2\pi e}{J[\delta]} \right) \quad , \tag{5.13}$$

it follows from Stam's inequality Eq. (5.10) that $D_0 \geq 0$, and that $D_0 = 0$ if and only if δ is Gaussian distributed. Conceptually, D_0 quantifies the non-Gaussianity of the distribution. With these notations, we can rewrite the mutual information Eq. (5.12) as

$$I[\tilde{\theta}, m] = H[m] - \int d\tilde{\theta} p(\tilde{\theta}) \left(\frac{1}{2} \ln(\frac{2\pi e}{J[\delta]}) + D_0 \right) . \tag{5.14}$$

Because $J[\delta] = J(\tilde{\theta})$ (see Section 5.3.1) we replace $J[\delta]$ with $J(\tilde{\theta})$ in Eq. (5.14) and obtain

$$\begin{aligned} I[\tilde{\theta}, m] &= H[m] - \int d\tilde{\theta} p(\tilde{\theta}) \left(\frac{1}{2} \ln(\frac{2\pi e}{J(\tilde{\theta})}) + D_0 \right) \\ &= H[m] - \int d\tilde{\theta} p(\tilde{\theta}) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\tilde{\theta})} \right) - \int d\tilde{\theta} p(\tilde{\theta}) D_0 \\ &= (H[m] - H[\tilde{\theta}]) + H[\tilde{\theta}] - \int d\tilde{\theta} p(\tilde{\theta}) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\tilde{\theta})} \right) - D_0 \ . \end{aligned} \tag{5.15}$$

We can verify that

$$H[\tilde{\theta}] - \int d\tilde{\theta} p(\tilde{\theta}) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\tilde{\theta})} \right) = H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right) \ . \tag{5.16}$$

Thus we can rewrite Eq. (5.15) as

$$\begin{aligned} I[\theta, m] &= I[\tilde{\theta}, m] \\ &= (H[m] - H[f(\theta)]) + H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right) - D_0 \ . \end{aligned} \tag{5.17}$$

Finally, with the definition of $C_0 = H[m] - H[f(\theta)]$ we arrive at the following expression for mutual information in terms of Fisher Information:

$$I[\theta, m] = \underbrace{H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right)}_{I_{\text{Fisher}}} + C_0 - D_0 \ . \tag{5.18}$$

Equation (5.18) is a general description of the relation between mutual information and Fisher information and is one of the main results of the paper. Its interpretation crucially depends on the magnitude of the two constants C_0 and D_0, as we will discuss in the following.

Lower bound on mutual information On one hand, if the noise is Gaussian, $D_0 = 0$. And adding additive noise can not decrease entropy, thus $C_0 \geq 0$. Therefore

$$I[\theta, m] \geq H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right) \ . \tag{5.19}$$

That is, if and only if the noise is Gaussian, I_{Fisher} is guaranteed to represent a lower bound on mutual information.

Upper bound on mutual information On the other hand, however, because Stam's inequality tells us that $D_0 \geq 0$, the first three terms on the *r.h.s* provide an upper bound on mutual information, thus

$$I[\theta, m] \leq H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right) + C_0 \ . \tag{5.20}$$

In particular, assuming vanishing noise, *i.e.* $H[\delta] \to 0$, we have $C_0 \to 0$. It follows that

$$I[\theta, m] \leq H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln \left(\frac{2\pi e}{J(\theta)} \right) \ . \tag{5.21}$$

As a result, I_{Fisher} represents a upper bound on mutual information in the small noise regime, which is the **opposite** of what [27] postulated.

Our analysis paints a more nuanced picture of when and how Fisher information can serve as a proxy for mutual information. In general, whether I_{Fisher} represents an upper or lower bound on mutual information is determined by the relative magnitudes of C_0 and D_0. The terms C_0 and D_0 quantify two different aspect of the noise: C_0 is related to the magnitude of the noise, while D_0 is related to the shape of the noise. It is important to note that even when the noise vanishes, *i.e.*, $C_0 \to 0$, D_0 can still be large and positive. Thus, assuming an asymptotic noise regime (*e.g.* large number of independent neurons) does *not guarantee* that I_{Fisher} asymptoti-

cally approximates mutual information. Only if both C_0 and D_0 are small, Fisher information is guaranteed to be a good proxy for mutual information.

5.4 Implications for neural coding models

In the following we discuss the implications of our main result to models of neural coding. Our derivation above was based on a standard input-output model where we assumed that both, the input and the output variable are one-dimensional. Neural coding models, however, frequently address the case where a one-dimensional stimulus variable θ is represented in the activity vector R of a population of noisy neurons. Technically, computing the mutual information $I[\theta, R]$ is not precisely the same problem that we have addressed above because R is high-dimensional. However, we can approximate the problem by formulating mutual information for a quantity that is the projection from R back to the stimulus space. Denoting the image of such projection as m, we consider the quantity $I[\theta, m]$ a surrogate of $I[\theta, R]$. Although it is likely that the projection results in some loss of information, the model in Eq. (5.6) can still provide a good and tractable approximation of the more complicated neural population coding model. This is particularly true if the projection is such that m preserves most of the information in R about θ, and the noise of the projection is approximately additive. The noise in m can be thought of as the "effective noise" which summarizes the noise characteristics of the whole neural population with regard to the stimulus dimension [144]. Note that

various studies have used similar surrogate formulations of mutual information in terms of projected quantities, by considering m as a particular estimator of θ (e.g. [17, 144, 27]).

5.4.1 Information measures of neural codes

Our theoretical predictions based on the input-output model are supported by recent numerical results. [185] performed systematic numerical measurements of the mutual as well as the Fisher information in the response of a population of neurons with bell shape tuning curves. They systematically varied population size and levels of response variability. Figure 5.1 depicts one of their simulation results for Gaussian neural noise with different Fano factors and different population sizes. The results show that under these conditions I_{Fisher} consistently *overestimates* mutual information. This supports our finding that Fisher information generally does not provide a lower bound on mutual information. Only as the population size and/or the integration time t increases, the effective noise becomes small but also more Gaussian, and Fisher information serves as an accurate proxy for mutual information.

In general, our analysis suggests that in order to use I_{Fisher} (as in Eq. (5.5)) as a proxy for the amount of information conveyed in a neural code, one should first examine the underlying noise characteristics before drawing any conclusions. In the continuous case, deviations from Gaussianity will result in the overestimation

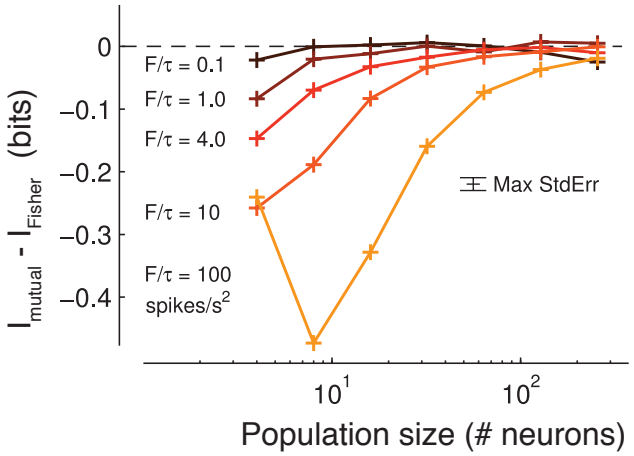

Figure 5.1: *Fisher information generally overestimates mutual information* (replotted from [185]). Shown is the difference between the numerically computed mutual information $I[\theta, R]$ and the Fisher information I_{Fisher} (as in Eq. (5.5) on the simulated response of a population of neurons. The spike counts of individual neurons was assumed to follow a Gaussian truncated at zero (to ensure positive values), with variance that follows the mean with a scale factor (Fano-factor F). Note, that the "effective noise" of the population is generally not Gaussian (see main text). Individual curves show the difference between mutual and Fisher information as a function of the population size, for different ratios between Fano factor and integration time τ. As the population size is small and the effective noise is large and non-Gaussian, Fisher information can significantly overestimate mutual information.

of mutual information. Or put in other words: I_{Fisher} by itself is not a good approximation for the mutual information in the neural code (even in the small noise regime), unless it is known that the noise characteristics is close to Gaussian.

This is especially important when studying neural codes based on multi-modal tuning curves, *e.g.* grid cells [86]. Fisher Information has been a popular quantity to analyze the code of the grid cell system. For example, it has been argued that the grid cell code has exponentially large capacity because Fisher information can grow exponentially with the number of neurons [162, 126]. Our results suggest that conclusions on coding capacity based on measures of Fisher information may be misleading. Fisher information of a neural code can be arbitrarily large without changing its mutual information if the effective noise in the neural representation is very non-Gaussian. Importantly, the mismatch between Fisher and mutual information can remain even when assuming large populations or vanishing noise. Neurons that exhibit multi-modal tuning curves are not uncommon and, besides grid cells, also include *e.g.*, disparity tuned neurons in primary visual cortex [43, 70] or ITD tuned neurons in owls [31]. Measures of coding efficiency for these neural systems are subject to similar concerns (see also [15]). With our derivation Eq. (5.18) we provide a way to precisely quantify the amount of overestimation of the mutual information.

5.5 Efficient coding interpretation

Efficient coding models are often formulated with regard to maximizing mutual information [10, 121]. In the following we show how we can rewrite Eq. (5.18) such that it provides an intuitive interpretation of the efficiency of a code. Specifically, we exploit the fact that

$$\begin{aligned} H[\theta] - \int d\theta p(\theta) \frac{1}{2} \ln(\frac{2\pi e}{J(\theta)}) &= \int d\theta p(\theta) \left(-\ln p(\theta) - \frac{1}{2} \ln(\frac{2\pi e}{J(\theta)}) \right) \\ &= \ln \left(\frac{\int \sqrt{J(\theta)} d\theta}{\sqrt{2\pi e}} \right) - \int d\theta p(\theta) \ln \left(\frac{p(\theta)}{\frac{1}{S}\sqrt{J(\theta)}} \right) \\ &= \ln \left(\frac{\int \sqrt{J(\theta)} d\theta}{\sqrt{2\pi e}} \right) - KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)}), \quad (5.22) \end{aligned}$$

where $KL(.)$ is the Kullback-Leibler divergence between two probability distributions [110]. The normalization constant S ensures that $\frac{1}{S}\sqrt{J(\theta)}$ is a proper probability density, *i.e.*, $S = \int \sqrt{J(\theta)} d\theta$.

With Eq. (5.22) we can rewrite the expression for mutual information Eq. (5.18) in terms of four meaningful and intuitive components:

$$I[\theta, m] = \frac{1}{2} \ln \left(\frac{S^2}{2\pi e} \right) - KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)}) - D_0 + C_0, \quad (5.23)$$

The first term can be interpreted as the overall coding resources available (in units of Fisher information). The second term $KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)})$ characterizes the information loss due to the mismatch between the input distribution and the way the

coding resources are distributed. The third term D_0 evaluates the information loss due to non-Gaussianity of the "effective noise". And the forth term C_0 is related to the overall level (entropy) of the noise.

5.5.1 Maximizing mutual information

An Efficient coding problem must be formulated with respect to a particular objective function and a set of constraints. We consider mutual information $I[\theta, m]$ the objective function to be maximized. The optimization is subject to a constraint on the overall resource budget available. We consider the total Fisher information $S = \int \sqrt{J(\theta)} d\theta$ to express this budget. S could be interpreted as the capacity of the code, because it is proportional to the number of discriminable states, e.g. the total discriminability, which is independent of the input distribution.

Vanishing noise regime Let's first consider a regime where the noise is vanishing, *i.e.*, the entropy of the noise $H[\delta] \to 0$. Many Efficient coding theories have been developed assuming such a regime (see *e.g.*,[27, 77]). Because $C_0 = H[m] - H[f(\theta)] \to 0$,

$$I[\theta, m] = \frac{1}{2} \ln \left(\frac{S^2}{2\pi e} \right) - KL(p(\theta) || \frac{1}{S} \sqrt{J(\theta)}) - D_0 \tag{5.24}$$

To maximize $I[\theta, m]$ with respect to the constraint on S, it is necessary that both $KL(p(\theta)||\frac{1}{S}\sqrt{J(\theta)})$ and D_0 are zero. This provides two necessary conditions for an efficient neural code.

First, in order to minimize the KL divergence $p(\theta) = \frac{1}{S}\sqrt{J(\theta)}$, *i.e.*, the neural system should distribute its total available coding resources according to the input distribution by choosing the appropriate transfer function f. It can be viewed as a probabilistic re-formulation of histogram equalization [111] with the important difference that what is equalized is not firing rates of neurons but rather the square-root of Fisher information. Using Fisher information has the advantage that we can formulate Efficient coding solutions without being limited to specific neural coding characteristics (tuning curves). Second, in order to minimize D_0 an efficient neural representation should exhibit an "effective" noise characteristics that is as close as possible to Gaussian.

Non-vanishing noise regime If the noise is large C_0 is non-zero. However, if C_0 does not depend on any of the other terms on the *r.h.s.* of Eq. (5.23) then, again, an efficient representation is one whose Fisher information (square-root) matches the input distribution and whose noise is Gaussian. Generally, we find that the form of the transfer function f may slightly change the difference between $H[m]$ and $H[f(\theta)]$, and therefore C_0, because of boundary effects induced by the limited output space. The dependence, however, is typically weak for noise that is not very large.

5.5.2 Signatures of Efficient coding

From above discussion we can identify two characteristic signatures of a system that efficiently encodes sensory information by maximizing mutual information between stimulus and representation. The first is that the system's coding resources are allocated such that

$$p(\theta) \propto \sqrt{J(\theta)} \ . \tag{5.25}$$

This simple relation is reminiscent of the optimal input distribution for a given noise channel in statistics [33, 145]). With this signature, we can probe the Efficient coding hypothesis by *e.g.* computing the Fisher information of an entire neural population based on electrophysiological measurements and then compare this distribution to the input (stimulus) distribution.

Previous work has mainly focused on characterizing Efficient coding in terms of neural tuning characteristics *e.g.*,[111, 88, 144, 181, 77]. For example, [77] have directly optimized the lower bound on mutual information proposed by [27] for a continuous parametric description of the population tuning characteristics. They found that the neural density, *i.e.*, the distribution of the neurons' preferred tuning values, should match the stimulus distribution $p(\theta)$. While under certain noise assumptions this is equivalent to our above proposed signature Eq. (5.25), it generally is not. Figure 5.3 demonstrates how two neural populations with the same number of neurons can have very different overall tuning characteristics yet still have the same Fisher information. While the neural density of one population (Fig. 5.2a)

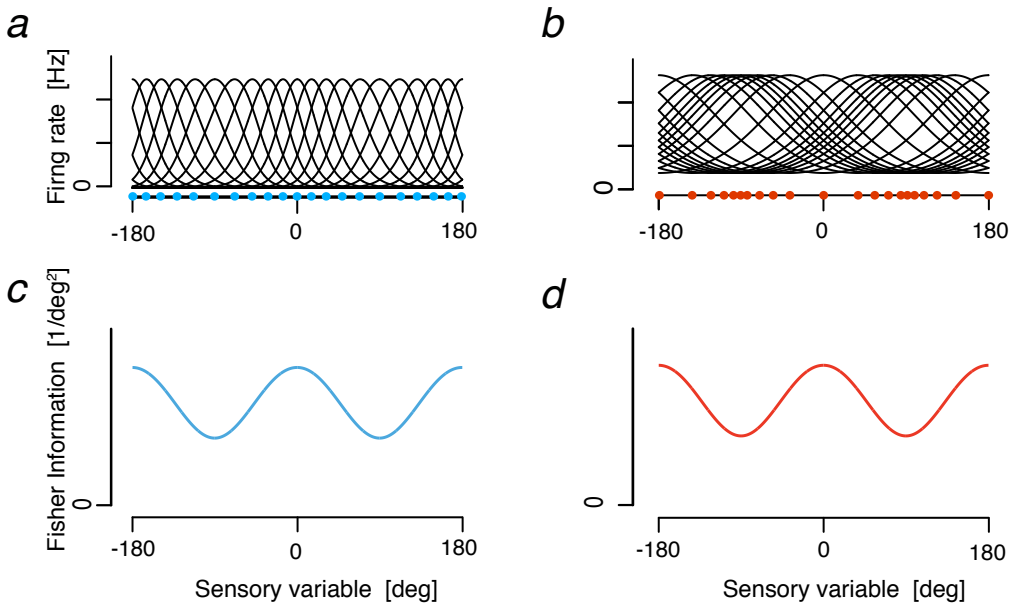

Figure 5.2: *Different population tuning solutions lead to equivalent distributions of Fisher information.* a,b) The tuning curves of two neural populations with the same number of neurons encoding a sensory (circular) variable. Each neuron's preferred tuning is indicated with a dot. The neural densities for both populations are different; the density for the population in a) is highest around zero while the density of the population in b) has peaks at $\pm$ 90 [deg]. c,d) Despite their differences in the distribution of the neurons, the Fisher information is identical for the two populations (up to a scale factor) suggesting that both neural populations are Efficient coding solutions for the same stimulus distribution (which is equivalent to the square-root of the Fisher information).

matches Fisher information (and thus the stimulus distribution [77]), the other population shows the opposite tuning characteristic where the peaks of the neural density coincide with the locations of minimal Fisher information (Fig. 5.2b).

This may explain why some of the known neural density distributions for sensory variables match the encoding accuracy of these variables (*e.g.*, orientation tuned neurons and spatial frequency tuned neurons in primary visual areas) while this is not the case for others variables. One example is the neural representation of heading direction. Neurons in area MST of the Macaque monkey are tuned for heading directions. Fig. 5.3a shows the histogram over the measured preferred heading directions of a large pool of MST neurons (replotted from [85]). More neurons are tuned to lateral directions while the population Fisher information (Fig. 5.3b) is maximal for forward and backward heading directions. This does not match the notion of an efficient neural population as proposed by [76] yet is fully consistent with our formulation Eq. (5.25) as demonstrated in Fig. 5.2b. Although measurements of the distribution of heading directions for a behaving primate do not exist, we can predict this distribution based on the measured Fisher information of the MST neural population as shown in Fig. 5.3c. Other examples demonstrating a mismatch between Fisher information and neural population density have also been reported by [69, 88].

Furthermore, we can establish a direct link between Fisher information $J(\theta)$, the stimulus (input) distribution $p(\theta)$, and perceptual discrimination threshold $d(\theta)$,

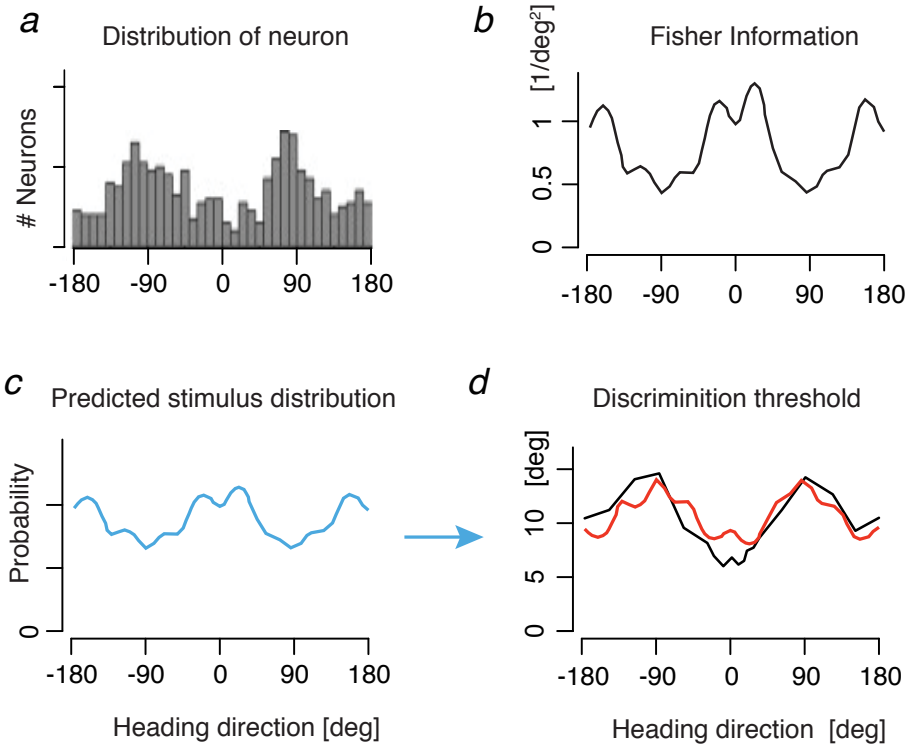

Figure 5.3: *Encoding of heading direction by neurons in area MST of the Macaque.* a) Distribution of measured preferred heading directions of MST neurons. More neurons are tuned for lateral rather than back and forth directions. b) Population Fisher information, however, shows peaks at forward and backward heading directions indicating that those directions are most accurately encoded. c) Based on our signature of Efficient coding (Eq. (5.25)) we predict the distribution of heading direction to follow the square-root of Fisher information (b). While exact statistical measurements for the distribution of heading direction are missing the prediction suggests that forward and backward headings are most frequent, which seems to be in agreement with everyday observations. d) Psychophysically measured discrimination thresholds for heading direction (black curve) nicely reflect the predicted discriminability based on the stimulus distribution (red curve), which represents our second signature. Data in a), b), and d) are replotted from [85] (Vestibular signals only).

independent of the tuning characteristics of the underlying neural representation. It has been shown that the inverse of the square root of Fisher information provides a lower bound on the discrimination threshold for unbiased [152] or biased [151] estimators. The bound is tight if the noise is Gaussian, which is one of the conditions we identified to maximize mutual information. Thus, if the encoding is maximally efficient the discrimination threshold $d(\theta)$ is determined as

$$d(\theta) \propto \frac{1}{p(\theta)} \ . \tag{5.26}$$

This represents a second signature for an efficient sensory representation: The discrimination threshold should directly match the inverse of the stimulus distribution (see [76]). As shown in Fig. 5.3d, this prediction is well matched with measured discrimination thresholds for heading direction. Again, we expect this relation to be hold irrespective of the specific tuning characteristics. For example, the density of the neurons may not follow the input distribution, but the discrimination threshold should.

5.6 Discussion

We have revisited and clarified the relation between Fisher information and mutual information in the context of neural coding. We derived a new result that describes a more general connection between Fisher and mutual information. In

particular, we demonstrated that Fisher information typically *does not represent a lower bound* on mutual information as frequently assumed based on the derivation by [27]. Rather, we found that the relation between mutual information and Fisher information (*i.e.*, I_{Fisher}; Eq. (5.5)) appears to be more nuanced. Fisher information is deeply linked to mutual information but generally does not provide a lower bound. It can be a lower bound but only if the noise is Gaussian. Furthermore, in the vanishing (small) noise regime Fisher information actually provides an *upper bound* on mutual information. If the noise is both Gaussian and small, then Fisher information provides a tight bound for mutual information. Thus, it is very important to note that assuming vanishing noise (asymptotic small noise limit) does not guarantee that mutual information is well approximated by Fisher information. The key coding characteristics that determines by how much Fisher information overestimates mutual information is the degree to which the noise is Gaussian. Previous numerical analyses support our result, showing that I_{Fisher} is generally larger than mutual information [185], and that Fisher information of a single neuron can be made arbitrary high without changing the entropy [15].

Our revised derivation of the relation between Fisher information and mutual information has significant impact the assessment of neural code and the theories of Efficient coding.If indeed I_{Fisher} generally provided a lower bound on mutual information, measures that increased Fisher information of a neural code would automatically imply an increase in coding efficiency. This could lead to an incorrect

assessment of how to increase coding efficiency , such as *e.g.*, by employing a code that uses multi-modal tuning curves (see grid cells). However, because we found that Fisher information (in the small noise regime) generally provides an upper bound on mutual information, the focus in assessing coding efficiency should be shifted towards measures that actually make the bound tight (rather than measures that incorrectly assume that the bound could be raised). This is an important conceptual difference. Our new formulation Eq. (5.23) allows us to quantify the difference between mutual information and Fisher Information, and to determine the precise conditions under which the two quantities becomes equivalent. We found two necessary conditions: First, the square root of Fisher information has to match the input distribution. And second, the effective noise should be Gaussian distributed.

Finally, our results may provide an explanation for some of the reported differences in the coding strategies of biological neural systems. Formulated with regard to Fisher information, we can specify multiple equivalent Efficient coding solutions for neural representations that are severely different in terms of their underlying neural tuning characteristics such as *e.g.*, in their neural density. This allows us to explain correlations between neural density and stimulus distribution that are inconsistent with previously proposed theories of Efficient coding that were directly formulated at the level of the neural tuning characteristics [76, 181, 77]. An exciting line of further research will be to understand in detail what additional constraints

favor one solution over the others.

Chapter 6

General discussions

6.1 Summary of the contributions

I have investigated how the idea of *efficient computation* can guide us understand certain aspects of the computations in the brain. In Chapter 3, I have demonstrated that such idea leads to a well-constrained yet powerful framework for understudying human perceptual behaviors (*efficient* both in term of *encoding* and *decoding*) that explains puzzling human performance in various psychophysical experiments involving simple perceptual decisions. Although I have only presented results for two visual stimulus variable, the general framework is readily applicable to other perceptual variables. Note that this framework also offers a principle way to address the common criticism of Bayesian models which argues that Bayesian models are lack of constraints. In Chapter 4, I demonstrate that the idea of efficiency,

coupled with a few reasonable assumptions, permits quantitative predictions on the functional architecture of the grid cell system in rodents. Such zero-parameter predictions beautifully match the data collected in recent neurophysiological experiments. These results are particularly striking in that these predictions, guided by first principle, are made before the key neurophysiological data [164] are published. It is a demonstration that theory in neuroscience, when formulated properly, could step ahead of the experiments. This study suggests that achieving efficiency of the neural computation as a fundamental design principle for neural circuits involving high-level cognition (i.e. representation of space). In Chapter 5, I analytically derive a general connection between mutual information and Fisher information. These two quantities are important both theoretically and practically. This clarifies an important theoretical issue which has shared some confusion recently in the neural coding literature. Additionallty, it also provides some powerful signatures of Efficient coding. These results may help guide future experimental tests of Efficient coding. Together, the results presented in this thesis support the idea that a common design principle, i.e. *achieving efficient computation*, may generalized across circuits processing both low-level and high-level cognitive functions.

6.2 Future directions

6.2.1 Efficient coding

Although Efficient coding [4, 10] has served as a major hypothesis for studying neural codes over the past few decades, its rigorous empirical tests have been generally difficult except for a few simple situations [3, 111, 46, 8, 120, 157, 155]. In Chapter 5, I derived a particular prediction for Efficient coding which states that the square root of Fisher information should match the input distribution. As a remark, this prediction is reminiscent of the results in statistics which suggest that the Jeffreys prior [97] achieves the channel capacity in the asymptotic limit [33, 34]. Also note that this prediction is quite different from redundancy reduction proposed by Horace Barlow[10]. These predictions are not mutually exclusive, rather they emphasize different aspects of Efficient coding. Redundancy reduction concerns the statistical relationship between the response of different output units (typically pairwise relationship), while the matching between Fisher information and input distribution concern about the relationship between the input and the response properties of output units when putting all together.

This prediction is feasible to test, because both the input distribution and Fisher information are experimentally measurable quantities in certain situations. The input distribution could be measured by the computing the summary statistics of natural scenes. In term of Fisher information, one way to measure it is to compute

the population Fisher information of a appropriate neural population measured neurophysiologically. However, several technical challenges remain with this approach. First, one has to find the appropriate neural population which represents the target stimulus dimension. Second, the sampling of the neurons has to be decent to guarantee that the population Fisher information computed from the sampled neurons is representative of the whole population. Third, computing the Fisher information requires knowing the statistical dependence between the neurons' response. Practically, this may be avoid by assuming the noise independence of the neurons [85]. However, the results obtained under such assumption have to be interpreted with caution. An alternative way to measure the Fisher information of the system is to probe the limit of the discriminability along a particular stimulus dimension. Such techniques are well-established in psychophysics [84]. Theoretically, it is known that Fisher information provides a bound for discriminability, and under certain conditions the bound is tight [152, 151].

In deriving the prediction on the matching between Fisher information and input distribution, I have relied on a constraint on Fisher information in the Efficient coding formulation. I consider this constraint to be a natural one under the proposed formulation, because it is the constraint that is invariant with respect to the re-parameterization of the stimulus variable. However, I should point out that constraints on other quantities, rather than Fisher information, are possible. For example, one could made constraints on the number of neurons, the average firing

rate, the max firing rate or the number of spike. The relationships between these different constrained remain to be investigated in future.

Furthermore, the matching between Fisher information and the input distribution is just one necessary condition, but not a sufficient condition, for Efficient coding. There could be more necessary conditions given more constraints. For example, one can put the number of spikes as another constrained, and further constrained the possible shape of the tuning curves. Mathematically, as the constraints become strong enough, one could in principle derive sufficient and necessary conditions (e.g. see [76]). In this situation, the optimal configuration may be unique. However, given the diversity of the organization of the population tuning curves as discussed in Chapter 5, it seems unlikely that a unique optimal tuning configuration is possible to explain the neural representation of all stimulus variables encoded by the brain. One particularly interesting future research direction would be study the optimal tuning configuration under different neurally realistic constraints.

6.2.2 Bayesian computation

Several important issues related to Bayesian computation presented in Chapter 3 remain to be resolved. The first involves how Bayesian computation could be implemented using biologically-plausible operations. In one study which is not included in this thesis, Wei & Stocker (2012) demontrates that a biologically plausible population-vector-like readout can approximate Bayesian decoder assuming a

particular efficient neural population [183]. Similar results have been derived independently by Ganguli & Simoncelli(2014) [77]. While these results suggested that possibilities that Bayesian computation could be performed using biologically plausible operations, it has certain limitations. One major limitation is that currently such results only hold when assuming a particular efficient neural population code, i.e. a wrapped neural population based on a homogeneous population. As we discussed in Chapter 3 and Chapter 5, there may be many different configurations of efficient codes. It is unclear whether these results on readout generalize to these scenarios or not.

Second, I have only considered one stage of Bayesian computation in a simple encoding-decoding cascade. However, as the neural computation consists many stages of processing, an promising conceptual idea is to model the process as hierarchical Bayesian inference [114]. One important future research question is to ask how Bayesian computation could be implemented using biologically plausible rules in this more general set-up. A particular challenge for deriving such implementation is to take into account of the combined effect of noise generated by each stage itself and the noise propagated from other stages.

Third, the Bayesian observer model proposed is based on an Efficient coding principle which maximizes the transmission of information [121]. As discussed earlier in Chapter 3 and Chapter 5, there are considerable experimental supports for such hypothesis in early sensory processing. However, it seems also reasonable to

consider an ideal observer based on an encoding model which minimize a risk with respect to a particular loss function. This is a meaningful problem if the goal of the system is to perform best for one particular task. For example, if the goal of is to minimize the mean squared error (L_2 loss) in an orientation estimation task, the encoding stage should be designed in a way such that the Bayes least squared estimator should exhibit the minimal Bayes risk among all the possible designs. In general, the optimal design of the neural code would be different from what derived from optimizing the mutual information. The question of precisely how these optimal neural code should be remains technically challenging. It is possible that one may need to rely on some numerical techniques to obtain the optima. However, once such optimal neural code is obtained, the Bayesian observer model would make predictions on the behaviors for the task. Such predictions could be directly compared to the psychophysical measurement to test the validity of the theory.

Fourth, so far I have only considered the perceptual biases and discrimination threshold in our observer model in Chapter 3 and Chapter 5. The bias and discrimination threshold roughly correspond to the first-order and the second-order statistics of the response distribution. A more powerful test of the observe model would be to consider the whole response distribution, which could be measured in certain estimation tasks (but in general not 2-AFC tasks). This may shed more light on the loss function that the perceptual system use.

6.2.3 Adaptation

In Chapter 3, I have proposed a Bayesian observer model which integrates the idea of Efficient coding and Bayesian decoding. One crucial assumption involved is that the prior belief is taken to reflect the long-term environmental statistics. This is a reasonable assumption if the environment is stationary. However, in general this should only be seen as a first-order approximation. There are scenarios when this assumption is likely to break down. For example, in typical perceptual adaptation experiments[80], the stimulus statistics also exhibit fluctuation in short-term time scale. In this case, it seems that the prior belief for the next stimulus should reflect such short-term input history as well. In general, it is reasonable to assume that the stimulus statistics at various time scales together determine the brain's prior belief for the next stimulus. However, to figure out precisely how the prior belief should be predicted from the stimulus history, one would need a mathematically rigorous formulation and more knowledge about the how these stimulus statistics fluctuate at various time scales in natural environments.

Once the connection between the stimulus history and the brain's prior belief for next stimulus event is established, applying a similar set of ideas presented in Chapter 3 would give an ideal observer model for behaviors during perceptual adaptation experiments. Previously, it has largely been a puzzle why perceptual adaptation would lead to various after-effects, including stimulus-specific change of discrimination threshold and repulsive perceptual bias [80, 141, 18, 118, 128, 142,

36]. The framework outline here may provide a normative account of these effects.

At the neural level, it is commonly observed that adaptation lead to various neural change, including gain change and tuning curve shift (for reviews, see [36, 158]). One interesting future direction would be to examine whether these changes of the neural responses could be explained by the principle of Efficient coding or not. Following the results in Chapter 3 and Chapter 5, the square root of the Fisher information should follow the prior belief in order to be efficient. Allowing for gain change and tuning curve shift, there are multiple solutions which all leads to the match between Fisher information and the prior. I speculate that this might provide a possible explanation for the diversity of the tuning curves changes observed in different experimental conditions [158]. In this view, the tuning curves as well as the stochasticity of the neurons change in order to adjust the Fisher information to match the prior, while whether the shift of individual tuning curve is attractive or repulsive is less crucial.

6.2.4 Grid cells

In Chapter 5, it is demonstrated that the idea of an efficient representation of space, i.e. representing space using the minimal number of neurons, quantitatively predicts the key functional architecture of the rodent grid cell system [86, 11, 164]. This investigation opens exciting future research directions.

First, it remains unclear how the orientation of the grids should be organized

and how the optimal grid should change with respect to the change of the shape of the environment. In the derivation in Chapter 5, it is essentially assumed that the 2-d environment is a disk. It would be interesting to see how the optimal grid should be change in a square shape environment. Although the change could be subtle, it may provide additional tests of the optimality of the grid code. Recent experimental observations indeed suggest that the grid orientations as well as the shape of the grids indeed depend on certain aspects of the geometry of the environment [165, 107]. It thus seems a natural future step to figure out whether the same principle, which predicts the scaling the grids, could also explain the alignment and deformation of the grids.

Furthermore, little is known in terms of how general the grid-like representation is beyond the representation of the animal's spacial location. In a recent study performed on monkey [102], it is reported that some neurons in EC exhibit grid pattern with respect to the visual space. Given these results, it is natural to ask whether a similar functional architecture as observed in rodent also exists for visual space. Last but not least, a more fundamental question to yet be asked is that for what kind of variables we should expect the brain forms a grid-like a representation. Answers to these questions would be helpful for better understanding of the designing principles of the neural representation in general.

Bibliography

[1] S-I Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[2] M. I. Anderson and K. J. Jeffery. Heterogeneous modulation of place cell firing by changes in context. *The Journal of Neuroscience*, 23(26):8827–8835, 2003.

[3] J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

[4] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183, 1954.

[5] B. B. Averbeck, P. E. Latham, and A. Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, 2006.

[6] V. Balasubramanian and M. J. Berry. A test of metabolically efficient coding in the retina. *Network: Computation in Neural Systems*, 13(4):531–552, 2002.

[7] V. Balasubramanian, D. Kimber, and M. J. Berry II. Metabolically efficient information processing. *Neural Computation*, 13(4):799–815, 2001.

[8] V. Balasubramanian and P. Sterling. Receptive fields and functional architecture in the retina. *The Journal of Physiology*, 587(12):2753–2767, 2009.

[9] H. B. Barlow. A theory about the functional role and synaptic mechanism of visual after-effects. *Vision: Coding and efficiency*, 363375, 1990.

[10] Horace B Barlow. Possible principles underlying the transformation of sensory messages. *Sensory communication*, pages 217–234, 1961.

[11] C. Barry, R. Hayman, N. Burgess, and K. J. Jeffery. Experience-dependent rescaling of entorhinal grids. *Nature Neuroscience*, 10(6):682–684, 2007.

[12] Mr. Bayes and Mr. Price. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions*, 53:370–418, 1763.

[13] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[14] J. M. Bernardo. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.

[15] M. Bethge, D. Rotermund, and K. Pawelzik. Optimal short-term population coding: when fisher information fails. *Neural Computation*, 14(10):2317–2351, 2002.

[16] W. Bialek. Thinking about the brain. In *Physics of bio-molecules and cells. Physique des biomolécules et des cellules*, pages 485–578. Springer, 2002.

[17] W. Bialek, F. Rieke, R. R. R. Van Steveninck, and D. Warland. Reading a neural code. *Science*, 252(5014):1854–1857, 1991.

[18] C. Blakemore, J. Nachmias, and P. Sutton. The perceived spatial frequency shift: evidence for frequency-selective neurones in the human brain. *The Journal of Physiology*, 210(3):727–750, 1970.

[19] C. N. Boccara, F. Sargolini, V. H. Thoresen, T. Solstad, M. P. Witter, E. I. Moser, and M.-B. Moser. Grid cells in pre-and parasubiculum. *Nature Neuroscience*, 13(8):987–994, 2010.

[20] T. Bonnevie, B. Dunn, M. Fyhn, T. Hafting, D. Derdikman, J. L. Kubie, Y. Roudi, E. I. Moser, and M.-B. Moser. Grid cells require excitatory drive from the hippocampus. *Nature Neuroscience*, 16(3):309–317, 2013.

[21] J. S. Bowers and C. J. Davis. Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389, 2012.

[22] S. E. Braun. Home range and activity patterns of the giant kangaroo rat, dipodomys ingens. *Journal of Mammalogy*, pages 1–12, 1985.

[23] J.B. Brayanov and M.A. Smith. Bayesian and "Anti-Bayesian" biases in sensory integration for action and perception in the size-weight illusion. *Journal of Neurophysiology*, 103:1518–1531, 2010.

[24] K. H. Britten, M. N. Shadlen, W. T. Newsome, and J. A. Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of Neuroscience*, 12(12):4745–4765, 1992.

[25] V. H Brun, M. K. Otnæss, S. Molden, H.-A. Steffenach, M. P. Witter, M.-B. Moser, and E. I. Moser. Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. *Science*, 296(5576):2243–2246, 2002.

[26] V. H. Brun, T. Solstad, K. B. Kjelstrup, M. Fyhn, M. P. Witter, E. I. Moser, and M.-B. Moser. Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex. *Hippocampus*, 18(12):1200–1212, 2008.

[27] N. Brunel and J.-P. Nadal. Mutual information, fisher information, and population coding. *Neural Computation*, 10(7):1731–1757, 1998.

[28] Y. Burak and I. Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS computational biology*, 5(2):e1000291, 2009.

[29] J. Burge and W. S. Geisler. Optimal defocus estimation in individual natural images. *Proc, Natl. Acad. Sci. USA*, 108(40):16849–16854, 2011.

[30] N. Burgess, C. Barry, and J. O'Keefe. An oscillatory interference model of grid cell firing. *Hippocampus*, 17(9):801–812, 2007.

[31] C. E. Carr and M. Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *The Journal of Neuroscience*, 10(10):3227–3246, 1990.

[32] M. Chalk, A. R. Seitz, and P. Series. Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, 10(2), 2010.

[33] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of bayes methods. *Information Theory, IEEE Transactions on*, 36(3):453–471, 1990.

[34] B. S. Clarke and A. R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994.

[35] D. D. Clarke and L. Sokoloff. Circulation and energy metabolism of the brain. *Basic neurochemistry: molecular, cellular and medical aspects*, 6:637–669, 1999.

[36] C. W. G. Clifford, M. A. Webster, G. B. Stanley, A. A. Stocker, A. Kohn, T. O. Sharpee, and O. Schwartz. Visual adaptation: neural, psychological and computational aspects. *Vision Research*, 47(25):3125–3131, 2007.

[37] D. M. Coppola, H. R. Purves, A. N. McCoy, and D. Purves. The distribution of oriented contours in the real world. *Proc, Natl. Acad. Sci. USA*, 95(7):4002–4006, 1998.

[38] R. Coultrip, R. Granger, and G. Lynch. A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks*, 5(1):47–54, 1992.

[39] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[40] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. CRC Press, 1979.

[41] H. Cramér. A contribution to the theory of statistical estimation. *Scandinavian Actuarial Journal*, 1946(1):85–94, 1946.

[42] B. T. Crane. Direction specific biases in human visual and vestibular heading perception. *PloS one*, 7(12):e51383, 2012.

[43] B. G. Cumming and A. J. Parker. Local disparity not perceived depth is signaled by binocular neurons in cortical area v1 of the macaque. *The Journal of Neuroscience*, 20(12):4758–4767, 2000.

[44] R.E. Curry. A Bayesian model for visual space perception. In *Seventh Annual Conference on Manual Control. NASA SP-281*, page 187ff, Washington D.C., 1972. NASA.

[45] L. F. Cuturi and P. R. MacNeilage. Systematic biases in human heading estimation. *PloS one*, 8(2):e56862, 2013.

[46] Y. Dan, J. J. Atick, and R. C. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *The Journal of Neuroscience*, 16(10):3351–3362, 1996.

[47] D. E. Davis, J. T. Emlen, and A. W. Stokes. Studies on home range in the brown rat. *Journal of Mammalogy*, pages 207–225, 1948.

[48] P. Dayan and L. F. Abbott. *Theoretical Neuroscience*. Cambridge, MA: MIT Press, 2001.

[49] L. de Almeida, M. Idiart, and J. E. Lisman. The input–output transformation of the hippocampal granule cells: from grid cells to place fields. *The Journal of Neuroscience*, 29(23):7504–7512, 2009.

[50] V. de Gardelle, S. Kouider, and J. Sackur. An oblique illusion modulated by visibility: Non-monotonic sensory integration in orientation processing. *Journal of Vision*, 10(10), 2010.

[51] A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *Information Theory, IEEE Transactions on*, 37(6):1501–1518, 1991.

[52] D. Derdikman and E. I. Moser. A manifold of spatial maps in the brain. *Trends in Cognitive Sciences*, 14(12):561–569, 2010.

[53] D. Derdikman, J. R. Whitlock, A. Tsao, M. Fyhn, T. Hafting, M.-B. Moser, and E. I. Moser. Fragmentation of grid cell maps in a multicompartment environment. *Nature Neuroscience*, 12(10):1325–1332, 2009.

[54] C. F. Doeller, C. Barry, and N. Burgess. Evidence for grid cells in a human memory network. *Nature*, 463(7281):657–661, 2010.

[55] E. Doi, J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner, T. A. Machado, L. H. Jepson, K. Mathieson, D. E. Gunning, A. M. Litke, L. Paninski, E. J. Chichilnisky, and E. P. Simoncelli. Efficient coding of spatial information in the primate retina. *The Journal of Neuroscience*, 32(46):16256–16264, 2012.

[56] C. L. Dolorfo and D. G. Amaral. Entorhinal cortex of the rat: topographic organization of the cells of origin of the perforant path projection to the dentate gyrus. *Journal of Comparative Neurology*, 398(1):25–48, 1998.

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

[58] A. W. F. Edwards. *Likelihood*. CUP Archive, 1984.

[59] A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. R. van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

[60] A. A. Faisal, L. P. J. Selen, and D. M. Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, 2008.

[61] Gustav Theodor Fechner. *Elemente der Psychophysik*. Breitkopf und Haertel, Leipzig, 1860.

[62] A. A. Fenton, H-Y Kao, S. A. Neymotin, A. Olypher, Y. Vayntrub, and N. Lytton, W. W .and Ludvig. Unmasking the ca1 ensemble place code by exposures to small and large environments: more place cells and multiple, irregularly arranged, and expanded place fields in the larger space. *The Journal of Neuroscience*, 28(44):11250–11262, 2008.

[63] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12):2379–2394, 1987.

[64] I. Fiete, Y. Burak, and T. Brookings. What grid cells convey about rat location. *The Journal of Neuroscience*, 28(27):6858–6871, 2008.

[65] B. J. Fischer. Bayesian estimates from heterogeneous population codes. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–7. IEEE, 2010.

[66] B. J. Fischer and J. L. Pena. Owl's behavior and neural representation predicted by Bayesian inference. *Nature Neuroscience*, 14(8):1061–1066, 08 2011.

[67] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pages 309–368, 1922.

[68] H. S. Fitch. Habits and economic relationships of the tulare kangaroo rat. *Journal of Mammalogy*, pages 5–35, 1948.

[69] D. C. Fitzpatrick, R. Batra, T. R. Stanford, and S. Kuwada. A neuronal population code for sound localization. *Nature*, 388(6645):871–874, 1997.

[70] D. J. Fleet, H. Wagner, and D. J. Heeger. Neural encoding of binocular disparity: energy models, position shifts and phase shifts. *Vision Research*, 36(12):1839–1857, 1996.

[71] B. R. Frieden. *Science from Fisher information: a unification*. Cambridge University Press, 2004.

[72] M. C. Fuhs and D. S. Touretzky. A spin glass model of path integration in rat medial entorhinal cortex. *The Journal of Neuroscience*, 26(16):4266–4276, 2006.

[73] M. Fyhn, T. Hafting, A. Treves, M.-B. Moser, and E. I. Moser. Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, 446(7132):190–194, 2007.

[74] M. Fyhn, T. Hafting, M. P. Witter, E. I. Moser, and M.-B. Moser. Grid cells in mice. *Hippocampus*, 18(12):1230–1238, 2008.

[75] M. Fyhn, S. Molden, M. P. Witter, E. I. Moser, and M.-B. Moser. Spatial representation in the entorhinal cortex. *Science*, 305(5688):1258–1264, 2004.

[76] D. Ganguli and E. P. Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. In *Advances in Neural Information Processing Systems*, pages 658–666, 2010.

[77] D. Ganguli and E.P. Simoncelli. Efficient sensory coding and bayesian decoding with neural populations. *Neural Computation*, 2014.

[78] W. S. Geisler, J. Najemnik, and A. D. Ing. Optimal stimulus encoders for natural tasks. *Journal of Vision*, 9(13), 2009.

[79] M. A. Georgeson and K. H. Ruddock. Spatial frequency analysis in early visual processing [and discussion]. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):11–22, 1980.

[80] J. J. Gibson and M. Radner. Adaptation, after-effect and contrast in the perception of tilted lines. i. quantitative studies. *Journal of Experimental Psychology*, 20(5):453, 1937.

[81] L. M. Giocomo, S. A. Hussaini, F. Zheng, E. R. Kandel, M.-B. Moser, and E. I. Moser. Grid cells use hcn1 channels for spatial scaling. *Cell*, 147(5):1159–1170, 2011.

[82] L. M. Giocomo, M.-B. Moser, and E. I. Moser. Computational models of grid cells. *Neuron*, 71(4):589–603, 2011.

[83] A. R. Girshick, M. S. Landy, and E. P. Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932, 2011.

[84] D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*, volume 1974. Wiley New York, 1966.

[85] Y. Gu, C. R. Fetsch, B. Adeyemo, G. C. DeAngelis, and D. E. Angelaki. Decoding of mstd population activity accounts for variations in the precision of heading perception. *Neuron*, 66(4):596–609, 2010.

[86] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

[87] J. B. Hales, M. I. Schlesinger, J. K. Leutgeb, L. R. Squire, S. Leutgeb, and R. E. Clark. Medial entorhinal cortex lesions only partially disrupt hippocampal place cells and hippocampus-dependent memory. *Cell Reports*, 9(3):893–901, 2014.

[88] N. S. Harper and D. McAlpine. Optimal neural population coding of an auditory spatial cue. *Nature*, 430(7000):682–686, 2004.

[89] M. E. Hasselmo, L. M. Giocomo, and E. A. Zilli. Grid cell firing may arise from interference of theta frequency membrane potential oscillations in single neurons. *Hippocampus*, 17(12):1252–1271, 2007.

[90] H. Helmholtz. *Treatise on Physiological Optics (transl.)*. Thoemmes Press, Bristol, U.K., 2000. Original publication 1867.

[91] G. H. Henry, B. Dreher, and P. O. Bishop. Orientation specificity of cells in cat striate cortex. *Journal of Neurophysiology*, 1974.

[92] A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian, and G. Tkačik. Variance predicts salience in central sensory processing. *eLife*, 3:e03722, 2014.

[93] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952.

[94] J. Jacobs, C. T Weidemann, J. F. Miller, A. Solway, J. F. Burke, X.-X. Wei, et al. Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, 16(9):1188–1190, 2013.

[95] J. Jastrow. Studies from the university of wisconsin: On the judgment of angles and positions of lines. *The American Journal of Psychology*, 5(2):214–248, 1892.

[96] M. Jazayeri and M. N. Shadlen. Temporal context calibrates interval timing. *Nature Neuroscience*, 13(8):1020–1026, 2010.

[97] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[98] M. Jogan and A.A. Stocker. Optimal integration of visual speed across different spatiotemporal frequency channels. In *Advances in Neural Information Processing Systems*, volume 26, pages 3201–3209, December 2013.

[99] M. Jones and B. C. Love. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34(04):169–188, 2011.

[100] Y. Karklin and M. S. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83–86, 2008.

[101] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[102] N. J. Killian, M. J. Jutras, and E. A. Buffalo. A map of visual space in the primate entorhinal cortex. *Nature*, 491(7426):761–764, 2012.

[103] D. C. Knill and W. Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

[104] K. Körding and D. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(15):244–247, January 2004.

[105] K. Körding and D. Wolpert. The loss function of sensorimotor learning. *Proc, Natl. Acad. Sci. USA*, 101(26):9839–9842, 2004.

[106] E. Kropff and A. Treves. The emergence of grid cells: Intelligent design or just adaptation? *Hippocampus*, 18(12):1256–1269, 2008.

[107] J. Krupic, M. Bauza, S. Burton, C. Barry, and J. O'Keefe. Grid cell symmetry is shaped by environmental geometry. *Nature*, 518(7538):232–235, 2015.

[108] J. Krupic, N. Burgess, and J. O'Keefe. Neural representations of location composed of spatially periodic bands. *Science*, 337(6096):853–857, 2012.

[109] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 5. California, 1951.

[110] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

[111] S. B. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch*, 36(910-912):51, 1981.

[112] S. B. Laughlin. Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, 11(4):475–480, 2001.

[113] S. B. Laughlin, R.R.R. van Steveninck, and J. C. Anderson. The metabolic cost of neural information. *Nature Neuroscience*, 1(1):36–41, 1998.

[114] T. S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448, 2003.

[115] E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer, 1998.

[116] P. Lennie. The cost of cortical computation. *Current Biology*, 13(6):493–497, 2003.

[117] S. Leutgeb, J. K. Leutgeb, C. A. Barnes, E. I. Moser, B. L. McNaughton, and M.-B. Moser. Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science*, 309(5734):619–623, 2005.

[118] E. Levinson and R. Sekuler. Adaptation alters perceived direction of motion. *Vision Research*, 16(7):779–IN7, 1976.

[119] W. B. Levy and R. A. Baxter. Energy efficient neural codes. *Neural Computation*, 8(3):531–543, 1996.

[120] M. S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[121] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[122] R. Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1(3):402–411, 1989.

[123] R. Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4(5):691–702, 1992.

[124] W. Maass. On the computational power of winner-take-all. *Neural Computation*, 12(11):2519–2535, 2000.

[125] A. Mathis, A. V. M. Herz, and M. B. Stemmler. Resolution of nested neuronal representations can be exponential in the number of neurons. *Physical Review Letters*, 109(1):018103, 2012.

[126] A. Mathis, A.V.M. Herz, and M. Stemmler. Optimal population codes for space: Grid cells outperform place cells. *Neural Computation*, 24(9):2280–2317, 2012.

[127] M. D. McDonnell and N. G. Stocks. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Physical Review Letters*, 101(5):058103, 2008.

[128] D. E. Mitchell and D. W. Muir. Does the tilt after-effect occur in the oblique meridian? *Vision Research*, 16(6):609–613, 1976.

[129] E. I. Moser, E. Kropff, and M.-B. Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.

[130] W. H. Mulders, M. J. West, and L. Slomianka. Neuron numbers in the presubiculum, parasubiculum, and entorhinal area of the rat. *Journal of Comparative Neurology*, 385(1):83–94, 1997.

[131] J. P. Nadal and N. Parga. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in neural systems*, 5(4):565–581, 1994.

[132] J. O'Keefe. Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, 51(1):78–109, 1976.

[133] J. O'Keefe and L. Nadel. *The hippocampus as a cognitive map*, volume 3. Clarendon Press Oxford, 1978.

[134] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:13, 1996.

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

[136] E. Pastalkova, V. Itskov, A. Amarasingham, and G. Buzsáki. Internally generated cell assembly sequences in the rat hippocampus. *Science*, 321(5894):1322–1327, 2008.

[137] T. Putzeys, M. Bethge, F. Wichmann, J. Wagemans, and R. Goris. A new perceptual bias reveals supoptimal population decoding of sensory responses. *PLoS Computational Biology*, 8(4):1–13, April 2012.

[138] G. J. Quirk, R. U. Muller, and J. L. Kubie. The firing of hippocampal place cells in the dark depends on the rat's recent experience. *The Journal of Neuroscience*, 10(6):2008–2017, 1990.

[139] C. R. Rao. Minimum variance and the estimation of several parameters. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 43, pages 280–283. Cambridge Univ Press, 1947.

[140] J. W. Strutt B. Rayleigh. *The theory of sound*, volume 2. Macmillan, 1896.

[141] D. Regan and K. I. Beverley. Spatial-frequency discrimination and detection: comparison of postadaptation thresholds. *JOSA*, 73(12):1684–1690, 1983.

[142] D. Regan and K. I. Beverley. Postadaptation orientation discrimination. *JOSA A*, 2(2):147–155, 1985.

[143] P. D. Rich, H-P Liaw, and A. K. Lee. Large environments reveal the statistical structure governing hippocampal representations. *Science*, 345(6198):814–817, 2014.

[144] F. Rieke, D.A. Bodnar, and W. Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 262(1365):259–265, 1995.

[145] J. Rissanen. Fisher information and stochastic complexity. *Information Theory, IEEE Transactions on*, 42(1):40–47, 1996.

[146] D. Rose and C. Blakemore. An analysis of orientation selectivity in the cat's visual cortex. *Experimental Brain Research*, 20(1):1–17, 1974.

[147] D. L. Ruderman. The statistics of natural images. *Network: Computation in neural systems*, 5(4):517–548, 1994.

[148] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 1994.

[149] E. Salinas. How behavioral constraints may determine optimal sensory representations. *PLoS Biology*, 4(12):e387, 11 2006.

[150] F. Sargolini, M. Fyhn, T. Hafting, B. L. McNaughton, M. P. Witter, M.-B. Moser, and E. I. Moser. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762, 2006.

[151] P. Seriès, A. A. Stocker, and E. P. Simoncelli. Is the homunculus "aware" of sensory adaptation? *Neural Computation*, 21(12):3271–3304, 2009.

[152] H.S. Seung and H. Sompolinsky. Simple models for reading neuronal population codes. *Proc, Natl. Acad. Sci. USA*, 90:10749–10753, November 1993.

[153] C. E. Shannon and W. Weaver. The mathematical theory of communication. *Urbana, Univ. of Illinois Press*, 1949.

[154] E. P. Simoncelli. Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2):144–149, 2003.

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

[156] N. A. Slade and R. K. Swihart. Home range indices for the hispid cotton rat (sigmodon hispidus) in northeastern kansas. *Journal of Mammalogy*, pages 580–590, 1983.

[157] E. C. Smith and M. S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, 2006.

[158] S. G. Solomon and A. Kohn. Moving sensory adaptation beyond suppressive effects in single neurons. *Current Biology*, 24(20):R1012–R1022, 2014.

[159] T. Solstad, C. N. Boccara, E. Kropff, M.-B. Moser, and E. I. Moser. Representation of geometric borders in the entorhinal cortex. *Science*, 322(5909):1865–1868, 2008.

[160] T. Solstad, E. I. Moser, and G. T. Einevoll. From grid cells to place cells: a mathematical model. *Hippocampus*, 16(12):1026–1031, 2006.

[161] H. Sompolinsky, H. Yoon, K. Kang, and M. Shamir. Population coding in neuronal systems with correlated noise. *Physical Review E*, 64(5):051904, 2001.

[162] S. Sreenivasan and I. Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature Neuroscience*, 14(10):1330–1337, 2011.

[163] A. J. Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959.

[164] H. Stensola, T. Stensola, T. Solstad, K. Frøland, M.-B. Moser, and E. I. Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72–78, 2012.

[165] T. Stensola, H. Stensola, M.-B. Moser, and E. I. Moser. Shearing-induced asymmetry in entorhinal grid cells. *Nature*, 518(7538):207–212, 2015.

[166] C. F. Stevens. Models are common; good theories are scarce. *Nature Neuroscience*, 3:1177–1177, 2000.

[167] O. Steward and S. A. Scoville. Cells of origin of entorhinal cortical afferents to the hippocampus and fascia dentata of the rat. *Journal of Comparative Neurology*, 169(3):347–370, 1976.

[168] L. F. Stickel and W. H. Stickel. A sigmodon and baiomys population in ungrazed and unburned texas prairie. *Journal of Mammalogy*, pages 141–150, 1949.

[169] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585, 2006.

[170] A. A. Stocker and E. P. Simoncelli. Sensory adaptation within a Bayesian framework for perception. *Advances in Neural Information Processing Systems*, 18:1289, 2006.

[171] E. Switkes, M. J. Mayer, and J. A. Sloan. Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis. *Vision Research*, 18(10):1393–1399, 1978.

[172] J. S. Taube, R. U. Muller, and J. B. Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. ii. effects of environmental manipulations. *The Journal of Neuroscience*, 10(2):436–447, 1990.

[173] A. Thue. Om nogle geometrisk-taltheoretiske theoremer. *Forh. Ved de skandinaviske naturforskeres*, pages 352–353, 1892.

[174] E. C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189, 1948.

[175] A. Tomassini, M. J. Morgan, and J. A. Solomon. Orientation uncertainty reduces perceived obliquity. *Vision Research*, 50(5):541–547, 2010.

[176] A. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

[177] T. Twer and D.I.A. MacLeod. Optimal nonlinear codes for the perception of natural colours. *Network: Computation in Neural Systems*, 12(3):395–407, 2001.

[178] R. Van den Berg, M. Vogel, K. Josic, and W.J. Ma. Optimal inference of sameness. *Proc, Natl. Acad. Sci. USA*, 109(8):3178–3183, 2012.

[179] J. H. Van Hateren. A theory of maximizing sensory information. *Biological Cybernetics*, 68(1):23–29, 1992.

[180] N. M. Van Strien, N. L. M. Cappaert, and M. P. Witter. The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. *Nature Reviews Neuroscience*, 10(4):272–282, 2009.

[181] Z. Wang, A.A. Stocker, and D.D. Lee. Optimal neural tuning curves for arbitrary stimulus distributions: Discrimax, Infomax and minimum L_p loss. In *Advances in Neural Information Processing Systems NIPS 25*, pages 2177–2185. MIT Press, December 2012.

[182] X.-X. Wei, J. Prentice, and V. Balasubramanian. The sense of place: grid cells in the brain and the transcendental number e. *arXiv preprint arXiv:1304.0031*, 2013.

[183] X.-X. Wei and A. A. Stocker. Bayesian inference with efficient neural population codes. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 523–530. Springer, 2012.

[184] X.-X. Wei and A.A. Stocker. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. In *Advances in Neural Information Processing Systems NIPS 25*, pages 1313–1321. MIT Press, December 2012.

[185] S. Yarrow, E. Challis, and P. Seriès. Fisher and shannon information in finite neural populations. *Neural Computation*, 24(7):1740–1780, 2012.

[186] M. M. Yartsev and N. Ulanovsky. Representation of three-dimensional space in the hippocampus of flying bats. *Science*, 340(6130):367–372, 2013.

[187] M. M. Yartsev, M. P. Witter, and N. Ulanovsky. Grid cells without theta oscillations in the entorhinal cortex of bats. *Nature*, 479(7371):103–107, 2011.

[188] E. Zohary, M. N. Shadlen, and W. T. Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 1994.