Efficient Coding and Bayesian Estimation with Neural Populations

by

Deep Ganguli

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Center for Neural Science
New York University
September 2012

Eero P. Simoncelli

Dedication

To my parents, Sham and Karobi Ganguli.

Acknowledgements

I would like to thank my advisor, Eero Simoncelli, for the support, enthusiasm, encouragement, and inspiration he has given me throughout the course of this research. It still amazes me how quickly he can arrive at solutions to my math problems (which mostly consist of symbols scribbled nervously on a chalkboard) by drawing simple pictures. I have learned so much from him.

I am grateful to members of the Simoncelli lab: Umesh Rajaeshekar, for teaching me everything I know about image processing and Matlab's implementation of the Fast Fourier Transform; Josh McDermott, for always answering my audio signal processing questions; Chaitue Ekhanadham, for being my resident calculus guru; Brett Vintch, for teaching me how to rock climb, learning how to snowboard with me, and always being down to ride bikes and chill at Le Basket; and Rob Young, for fixing my computer every time it broke and not being too mad when it was my fault.

I am especially grateful to Jeremy Freeman, who in many ways was my second advisor. His contributions include thoroughly reading and editing drafts of my papers, helping me put together talks, teaching me Adobe Illustrator and aspects of design so I could make better figures, talking me through statistical analyses, and selflessly championing my work to anyone who would listen. He is also a fantastic friend. I owe him one for introducing me to Kelly Gannon, whose contributions include (but are not limited too): delicious home cooked meals, bike adventures in Brooklyn, driving me to my first job interviews, encouraging me to never stop skating (even though it hurts more when I fall these days because I'm older now and the ground seems further away), and tons of love and support.

Abstract

The efficient coding hypothesis asserts that sensory systems evolved to maximize information transmitted to the brain about the environment. We develop a precise and testable form of this hypothesis in the context of encoding a sensory variable in the responses of a population of noisy neurons, each characterized by a tuning curve. We obtain a closed form solution for the information maximizing tuning curves as a function of the prior probability of sensory variables encountered in the environment. The solution states that more cells with narrower tuning widths should be allocated to encode higher probability stimuli. We extend our result to predict the discrimination performance of a perceptual system operating on the efficient neural representation, and find that the best achievable discrimination thresholds are inversely proportional to the sensory prior. The predicted relationships between empirically measured stimulus priors, physiological tuning properties, and perceptual discriminability are remarkably well matched to data obtained for two auditory and three visual variables. We also derive a novel decoder that performs Bayesian estimation by utilizing the prior information embedded in the preferred stimuli of the optimal tuning curves. Similar to the population vector, our decoder computes weighted averages of the preferred stimuli. However, the firing rates are not used directly as weights, but are first convolved with a linear filter then exponentiated. We map this simple cascade onto a compact, biologically plausible neural circuit. The results in this thesis provide a strong link between two dominant theories in sensory neuroscience − efficient coding and Bayesian estimation − and suggest how to relate both ideas directly to data from physiological and perceptual experiments.

Contents

Dedication . . . iii

Acknowledgements . . . iv

Abstract . . . v

List of Figures . . . ix

List of Tables . . . x

1 Introduction **1**

1.1 The efficient coding hypothesis . . . 2

1.2 Perception as Bayesian inference . . . 6

1.3 Thesis outline . . . 11

2 Implicit Encoding of Prior Probabilities in Optimal Neural Populations **14**

2.1 Introduction . . . 14

2.2 Encoding model and resource constraints . . . 16

2.3 Objective function . . . 17

2.4 How to optimize? . . . 19

2.4.1 Density and gain for a homogeneous population . . . 20

2.4.2 Density and gain for a heterogeneous population . . . 22

2.4.3 Objective function and solution for a heterogeneous population 24

2.5 Extensions 26

2.5.1 Monotonic tuning curves 26

2.5.2 Generalization to Poisson-like noise distributions 29

2.6 Discussion 32

3 Efficient Sensory Coding Predicts the Heterogeneities of Neural Populations and Perception 35

3.1 Introduction 35

3.2 Results 36

3.3 Discussion 41

3.4 Methods 44

3.4.1 Estimating environmental distributions 44

3.4.2 Cell density estimates 47

3.4.3 Estimating mutual information 48

3.4.4 Estimating normalized information 49

3.4.5 Testing the Fisher bound on mutual information 52

3.5 Supplementary information 53

3.5.1 Another orientation example 53

3.5.2 Predictions of gain 54

3.5.3 Predictions of alternative efficient coding frameworks 55

3.5.4 Predictions of an alternative optimality principle 58

4 Neural Implementation of Bayesian Estimation using Efficient Population Codes 68

4.1 Introduction 68

4.2 Efficient encoding model 73

4.3 Bayesian estimation 75

4.4 An importance sampling approximation to Bayesian estimation . . 78

4.5 The Bayesian population vector 79

4.6 Simulations 81

4.7 Experimental predictions 83

4.8 Discussion 85

5 General Discussion 93

Bibliography 98

List of Figures

2.1 Model Parameterization 21

3.1 Experimental predictions 62

3.1 Experimental predictions continued 63

3.2 Information transmitted 64

3.3 Deviations between estimated and predicted cell density 64

3.4 Another orientation example. 65

3.5 Gains 65

3.6 Comparison to ICA and sparse coding 66

3.7 Testing the discrimax predictions 67

4.1 Efficient encoding model 73

4.2 Decoding framework 76

4.3 Bayesian population vector 90

4.4 Optimality of the Bayesian population vector over the standard population vector 91

4.5 Bias variance properties of the Bayesian population vector 92

List of Tables

2.1 Closed form solution for optimal neural populations with unimodal tuning curves. 25

2.2 Closed form solution for optimal neural populations with monotonic tuning curves. 28

3.1 Normalized information 52

3.2 Testing the Fisher bound on mutual information 53

3.3 Comparison of efficient coding frameworks 56

Chapter 1

Introduction

As organisms interact with the world, their sensory systems are constantly bombarded with physical signals. In the visual system, photons impinging on photoreceptors in the retina give rise to patterns of spiking activity in neurons a few synapses away. In the auditory system, sound waves entering the ears cause a membrane to vibrate, which subsequently causes neurons to produce spikes. It is by virtue of these patterns of spiking activity that the brain is able to perceive the environment and allow an organism to produce actions and memories relevant for its survival.

A fundamental goal in sensory neuroscience is to understand the transformations of natural signals into neural representations that enable perception. Inspired by this goal, many investigators have carefully characterized the statistical properties of the natural environment, the response properties of neurons in early sensory systems, and the perceptual capabilities of human observers. These investigations have produced an enormous wealth of data about these three pieces of the puzzle. To understand the entire picture requires a theoretical framework that can illuminate connections between each piece, and be tested with the existing data. The goal of this thesis is to establish such a theoretical framework.

Two dominant theories in sensory neuroscience are explored in this thesis. The first theory, proposed by Barlow over 50 years ago, hypothesizes that early sensory systems evolved to maximize the amount of information transmitted to the brain about the environment, while minimizing metabolic costs [1,2]. This statement has been termed "the efficient coding hypothesis", and can be quantified within the framework of information theory [3]. The second theory is inspired by Helmholtz, who qualitatively described perception as a process of inference, in which human observers combine their noisy measurements and prior knowledge of the environment to construct an estimate of the physical world [4]. Bayesian statistics provides this notion with a quantitative grounding that can be used to develop optimality principles for perception. In the following sections we elaborate on these two theories, their existing support, and our new contributions (which include a previously unforeseen connection between the two theories).

1.1 The efficient coding hypothesis

Coding efficiency is a well known objective for the design of machine systems. It is rooted in the mathematics of information theory, which was initially developed to find fundamental limits on signal processing algorithms for the compression, transmission, and storage of data [3]. Remarkably, these ideas were also deemed relevant for perceptual and biological systems soon after [1,2]. These early papers argued that sensory and perceptual systems should exploit statistical regularities in the natural environment to efficiently encode sensory information. Although this hypothesis was developed over 50 years ago, it has been remarkably difficult to develop a quantitatively precise formulation of efficient coding that is simple to test with both physiological and perceptual data.

The efficiency of a given sensory representation depends on specifying four ingredients: (1) the family of possible neural transformations that dictate how a natural signal is encoded into neural activity; (2) the types of signals that are to be encoded, and their statistical properties; (3) the noise processes involved in the inputs and outputs of the system; and (4) the metabolic costs of building, operating, and maintaining the system. Given these ingredients, one can examine the statistical properties of environmental signals, and show that a transformation optimized according to a statistical optimality criterion, subject to the relevant constraints, provides a good description of the response properties of a set of sensory neurons [5]. The optimization of any formulation that realistically incorporates all the relevant ingredients is generally intractable. Nevertheless, several simplified formulations have achieved success at explaining the response properties of neurons in early sensory systems, and in some cases, the implications of coding efficiency for perception.

The simplest formulation of the efficient coding hypothesis considers a single neuron that is to encode scalar input with a continuous valued monotonic response function, which is bounded between a minimum and maximum response value. The neural response is assumed to be deterministic, and the frequency of occurrence of values of the input in the environment is modeled with a probability distribution. In this setup, the information maximizing neural response function is expressed as proportional to the cumulative probability distribution of the input. Intuitively, this solution allocates more sensitive regions of the neurons dynamic range to encode frequently occurring inputs with high fidelity, at the cost of encoding less frequent inputs with low fidelity. In a pioneering study, Laughlin measured the probability distribution of contrast levels in natural scenes, and found that the contrast response function of a fly large monopolar cell indeed closely resembled the cumulative of this distribution [6].

Inspired by this observation, several papers examined the optimal nonlinear transfer function in the presence of neural noise, and found that it could be analytically expressed as a power law function of the cumulative distribution, with the exponent depending on the precise noise properties and optimality criteria [7, 8]. These theoretical results were not directly compared to data. Further studies showed that some neurons dynamically update their monotonic response function to changes in stimulus statistics in order to maximize information transmission, [9, 10]. In all noiseless and noisy single neuron cases discussed so far, the optimal solution involves allocating greater response range (steeper portions of the transfer function) to more frequently occurring events.

Efficient coding formulations have also been developed to derive optimal coding schemes for populations of neurons. Some formulations model neurons as having linear receptive fields, scalar output non-linearities, and no output noise [11, 12]. For the case of no output noise, maximizing information is equivalent to minimizing the statistical dependency (or reducing the redundancy) between the outputs of all the neurons in the population [13]. Intuitively, this objective function asserts that an encoding scheme cannot be efficient if the effort of encoding any particular piece of information is duplicated in more than one neuron [5]. If the input distribution is considered to be a subset of natural images collected from the environment, and there is a power constraint on each neuron, a pioneering study found that the numerically optimized receptive fields are localized, oriented, and bandpass filters, which resemble the receptive fields of neurons in primary visual cortex (V1) [14]. Subsequent work compared the properties of these derived receptive fields to those of physiologically measured receptive fields, and found that there was good agreement between theory and data for subset of these properties [15].

Neurons are noisy (see [16] for review of noise in the nervous system) and, intuitively, some amount of redundancy in a neural code is desirable to combat the noise. Several variations of the efficient coding framework have attempted to incorporate both additive input noise and additive output noise into population coding models. For analytical tractability, these formulations assume Gaussian distributions for the inputs and both noise sources, and that a neural population can be described by a single linear receptive field that is convolved with the input. The optimal linear receptive field exhibits a center surround structure, which resembles the receptive fields of retinal ganglion cells [17–19], and can predict perceptual phenomena such as the contrast sensitivity function [18], and the oblique effect [20]. Recent work that relaxes the assumptions about convolutional receptive fields and Gaussian noise, shows that the information maximizing neural population also exhibits (heterogeneous) center-surround receptive fields with rectifying nonlinearities that are consistent with the detailed response properties of retinal ganglion cells [21].

An alternative view of efficient coding derives motivation from the steep metabolic costs of creating and maintaining a neural system [22–24]. Several papers argue for an “economy of spikes”, in which representational accuracy of a neural system should naturally trade off with the cost of firing an action potential. A quantitative formulation of this hypothesis [25] was shown to be consistent with the spiking response properties of retinal ganglion cells [26]. An alternative formulation, known as sparse coding, aims to derive receptive fields that allow for a minimum mean squared error linear reconstruction of an ensemble of inputs, subject to a sparsity constraint on the total number of active neurons [27,28]. The optimal receptive fields for encoding natural images resemble those of V1 simple cells [27]. For encoding natural sounds, the optimal receptive fields closely resemble those of auditory nerve fibers [28]. In the visual case, these receptive fields often exhibit response properties that are inconsistent with some aspects of the known physiology [29].

Several classical variants of the efficient coding hypothesis are supported to varying degrees by physiological or perceptual data. However, a formulation that provides a unifying, and easily testable description of the relationship between sensory statistics, neural coding, and perception is still lacking. Furthermore, although coding efficiency may be a reasonable objective for early stages of sensory processing, it seems unlikely to explain more specialized later stages responsible for producing actions that are directly relevant to the survival of an organism [30].

1.2 Perception as Bayesian inference

Organisms form percepts of the environment from the spiking activity of early sensory systems. But how can robust percepts of an uncertain world arise from noisy neural responses? Almost a century ago, Hermann von Helmholtz conjectured that sensory perception arises from the proper fusion of incoming sensory measurements with prior knowledge of the world [4]. Approximately 50 years later, ET Jaynes proposed a quantitatively precise formulation of this statement.

In an article submitted to the IEEE Transactions in Information Theory, Jaynes proposed that sensory transformations can be described with a calculus of probabilities known as Bayesian inference, and that the brain may be implementing this calculus [31]. Although the paper was rejected for being implausible and irrelevant at the time, Bayesian models of neural and perceptual systems have since become exceedingly popular [e.g., 32, 33].

Bayesian models of perception consist of three natural ingredients: (1) a likelihood function, which represents an observers degree of belief about an environmental signal based on sensory evidence; (2) a prior probability distribution, which characterizes an observers experience regarding the plausibility of different signal values in the environment; and (3) a loss function, which specifies the cost of making perceptual errors. Helmholtz's fusing of prior knowledge with current sensory information, corresponds to a multiplication of the prior distribution and likelihood function. This computation results (after normalization) in a *posterior* probability distribution, which is rationally biased towards the prior when the sensory evidence is weak, and discounts the prior when the sensory evidence is strong. A perceptual estimate can be obtained by choosing the signal value that minimizes the expectation, over the posterior distribution, of the loss function. Such an estimate is said to be optimal with respect to an observers prior, likelihood, and loss function. Optimal Bayesian estimation explains human performance in perceptual [34–36], sensorimotor, [37, 38], and cognitive [39] tasks remarkably well, and a number of review articles document progress to date [32, 40–44].

Specifying an observer's prior, likelihood and loss function for a particular task is challenging. If the observer has access to a prior distribution that was acquired from experience in the environment, then it may be possible to directly estimate the prior from the environment [36]. Such priors can also be estimated from psychophysical measurements of bias and discriminability, assuming an observer's loss function is known [35, 36, 45]. Specifying an observer's likelihood function requires knowledge of how the signal of interest is represented within the brain, including a characterization of physiological noise properties [46]. In perceptual experiments, investigators have incorporated noise into the external sensory inputs, which can provide insights into the properties of internal noise [47]. Specifying an observers loss function is difficult as subjects exhibit substantial variability in their behavior in an experimental situation involving rewards [e.g., 48, 49]. Nevertheless, given the experimental support for Bayesian models of perception, it seems natural to assume that the brain is implementing Bayesian estimation.

Specifying how the probability distributions required by the Bayesian machinery may be learned, represented, and computed within the brain poses a significant challenge. A number of authors suggest that probability distributions are explicitly represented in the firing rates of a neural population [e.g., 50–53]. In these formulations, operations such as computing the mean of a distribution reduce to a linear function of the firing rates, and probability distributions can be dynamically updated via gain changes. However, a detailed formulation of this sort needs to be verified with experimental data and address the calculation errors inherent to noisy representations of probability distributions.

Alternatively, many papers consider the *implicit* encoding of probabilities in the noisy responses of a population of neurons [e.g., 54–62]. These formulations are built around tuning curve descriptions of neurons, which are common in sensory neuroscience. A tuning curve characterizes a neuron's selectivity and average response rate for encoding scalar stimulus values. Assuming Poisson distributed neuronal variability, several papers show that a log-likelihood function can be expressed as a firing rate weighted sum of the log tuning curves in a population [57, 59, 61]. Further work shows that adding the firing rates of two neural populations, each responding to different sensory inputs, implicitly multiplies the likelihoods they represent [62]. This result provides a plausible neural calculus for behavioral evidence that suggests human observers take uncertainty into account when fusing multiple sources of information [42].

In the Bayesian framework, the likelihood function must be multiplied by the prior distribution in order to obtain a posterior distribution. Several proposals exist for how this may occur in the brain. For example, if a log-prior is encoded as a firing weighted sum of the log tuning curves in a neural population separate from the encoding population, then adding the spikes of the two populations implicitly computes a posterior distribution (analogous to the case of combining likelihoods) [62]. This proposal leads to a noisy representation of the prior (due to neuronal variability), but may be useful for encoding contextual or task dependent priors that must evolve over time. Alternative proposals suggest that *samples* from a prior distribution may be represented in spontaneous spiking activity, and stimulus evoked activity may represent samples from a posterior distribution [63, 64]. This framework is consistent with some physiological evidence [64], and can provide an algorithmic explanation for perceptual multi-stability (where a percept switches over time due to ambiguous sensory input) [65]. However, it is unclear how an explicit sensory estimate may be obtained from such neural responses, or how inference algorithms based on sampling strategies may be implemented in canonical neural circuitry.

Once a posterior distribution is represented in the brain, an optimal perceptual estimate may be derived according to a specified loss function. Determining how loss functions are represented in the brain, and dependent on reward contingencies, is an area of active research in the nascent interdisciplinary field of “neuro-economics” [66–68]. Because it is difficult to specify what loss function the brain may encode and employ for a perceptual task, particular loss functions are often chosen for gaining mathematical tractability to derive optimal Bayesian estimation strategies. For example, for a loss function that incurs zero penalty if the perceptual estimate is equal to the sensory input, and a large penalty for all other estimates, the optimal Bayesian estimate corresponds to computing the sensory input that maximizes the posterior distribution (a MAP estimate). Consider a downstream population implementing a MAP estimate from the neural responses of an encoding population, which is characterized by tuning curves and independent Poisson noise. The decoder must first explicitly compute a likelihood by computing a sum of tuning curves, each weighted by the firing rates of the encoding population, then add the result to a log prior distribution to obtain a log posterior distribution. The MAP estimate would then be the value associated with the decoder neuron with the maximum response, corresponding to a winner take all readout mechanism.

A simpler decoder − the population vector (or center of mass decoder) − produces a perceptual estimate by computing a firing rate weighted average of the preferred stimulus values of each neuron. The population vector has a rich history in experimental neuroscience, where it has been used to accurately predict the direction of arm movements from neural responses measured in a variety of motor and premotor areas [69–72], the orientation of visual stimuli from responses in primary visual cortex [73], the direction of saccades from responses in the superior colliculus [74], and the position of a rat from the responses of place cells in the hippocampus [75]. Recent theoretical work has shown that the population vector can compute the mean of a posterior distribution [76,77], which is optimal for minimizing a squared loss function, and is known as a Bayes least squares estimator. However, these results rely on strong assumptions about the encoding population, which are asserted for the explicit purpose equating a Bayes least squares estimator to a population vector.

Experimental evidence suggests that human judgments of many sensory attributes are consistent with optimal Bayesian estimation, in which noisy sensory measurements are combined with prior knowledge to obtain perceptual estimates. An important problem in sensory neuroscience is to develop a neurally plausible calculus for Bayesian computations that can be verified with physiological and perceptual data.

1.3 Thesis outline

In this thesis, we consider the efficient encoding of sensory information in neural populations, and the implications of coding efficiency for Bayesian theories of perception. In Chapter 2 we start by considering the encoding of a sensory variable with a heterogeneous population of noisy neurons, each responding selectively to a particular range of values. The accuracy with which a neural population encodes a stimulus value depends on the number of cells that respond to that value (cell density), their selectivity (tuning widths), and their response levels (gain). We optimize these parameters for a family of objective functions, that include maximizing mutual information (a variant of the efficient coding hypothesis) and maximizing perceptual discriminability, as special cases. The optimal solutions for the cell density, tuning width, and gain may be obtained in closed form, and are expressed as a power law function of the prior probability of stimulus values encountered in the environment. The exponent of the power law depends on the objective function. In addition, the optimal neural populations impose strong limitations on the ability of an organism to discriminate different values of the encoded variable. The minimum achievable discrimination thresholds are also expressed as power law functions of the prior, with the exponent depending on the objective function. As a result, our optimal coding framework predicts an explicit relationship between the statistical properties of the environment, the allocation and selectivity of neurons within populations, and perceptual discriminability.

In Chapter 3 we test these relationships for the efficient coding objective function in the context of two auditory attributes (acoustic frequency and modulation frequency), and three visual attributes (local orientation, spatial frequency, and retinal speed). The prior probability of each attribute is estimated from large databases of natural scenes and sounds. Physiological data are taken from single-cell electrophysiological recordings in primate or cat that report the independently measured tuning widths of a large population of neurons as a function of their stimulus preferences. These measurements were made in a diverse set of brain areas (the auditory nerve fibers, the inferior colliculus, the primary visual cortex, and the middle temporal cortex), each chosen based on a substantial literature identifying the tuning properties of those neurons for the stimulus feature of interest. Discrimination thresholds for each sensory attribute are obtained from human perceptual experiments. For all cases, we find that our efficient coding predictions of the relationship between the environment, physiology, and perception are well supported by the data.

In Chapter 4 we derive a novel decoder that can correctly extract the prior information embedded in the efficient population code, and combine it with likelihood information to produce a Bayesian least squares stimulus estimate. We map the decoder onto a biologically plausible cascade that consists of linear filtering, a static non-linearity, and divisive normalization. We show that the decoder closely approximates an omniscient Bayes least squares estimator that has full access to the prior and likelihood. Finally, we discuss how to test for signatures of this decoders use with psychophysical measurements of estimation bias.

Chapter 2

Implicit Encoding of Prior Probabilities in Optimal Neural Populations

2.1 Introduction

Many bottom up theories of neural encoding posit that sensory systems are optimized to represent sensory information [1, 2]. The optimality of a given sensory representation depends on specifying four components: (1) the family of possible neural transformations that dictate how a natural signal is encoded into neural activity; (2) the types of signals that are to be encoded, and their statistical properties; (3) the noise processes involved in the inputs and outputs of the system; and (4) the metabolic costs of building, operating, and maintaining the system. It is difficult to test whether a sensory system is optimal because the optimization of any formulation that attempts to correctly incorporate all of the relevant ingredients is generally intractable.

A substantial literature has considered simple population coding models in which each neuron's mean response to a scalar variable is characterized by a tuning curve [e.g., 54–62]. For these models, several papers have examined the optimization of Fisher information, which expresses a bound on the mean squared error of an unbiased estimator [78–81]. In these results, the distribution of sensory variables is assumed to be uniform and the populations are assumed to be homogeneous with regard to tuning curve shape, spacing, and amplitude.

The distribution of sensory variables encountered in the environment is often non-uniform, and it is thus of interest to understand how variations in probability affect the design of optimal populations. It would seem natural that a neural system should devote more resources to regions of sensory space that occur with higher probability, analogous to results in coding theory [82]. At the single neuron level, several publications describe solutions in which monotonic neural response functions allocate greater dynamic range to higher probability stimuli [6–8, 11]. At the population level, non-uniform allocations of neurons with identical tuning curves have been shown to be optimal for non-uniform stimulus distributions [83, 84].

Here, we examine the influence of a sensory prior on the optimal allocation of neurons and spikes in a population, and the implications of this optimal allocation for subsequent perception. Given a prior distribution over a scalar stimulus parameter, and a resource budget of N neurons with an average of R spikes/sec for the entire population, we seek the optimal shapes, positions, and amplitudes of tuning curves. We assume a population with Poisson-like spiking (which may include correlations), and consider a family of objective functions based on Fisher information. This family includes lower bounds on mutual information, and the minimum attainable perceptual discrimination performance as special cases. We then approximate the Fisher information in terms of two continuous resource variables, the density and gain of the tuning curves. This approximation allows us to obtain a closed form solution for the optimal population. For all objective functions, we find that the optimal tuning curve properties (cell density, tuning width, and gain) are power-law functions of the stimulus prior, with exponents dependent on the specific choice of objective function. Through the Fisher information, we also derive a bound on perceptual discriminability, again in the form a power-law of the stimulus prior. Thus, our framework provides direct and experimentally testable links between sensory priors, tuning properties of optimal neural representations, and perceptual discriminability.

2.2 Encoding model and resource constraints

We start with a conventional model for a population of N neurons responding to a single scalar variable, s [e.g., 54–62]. The number of spikes emitted (per unit time) by the nth neuron is a sample from an independent Poisson process, with mean rate determined by its tuning function, $h_n(s)$. The probability density of the population response can be written as

$$p(\vec{r}|s) = \prod_{n=1}^{N} \frac{h_n(s)^{r_n}\ e^{-h_n(s)}}{r_n!}. \tag{2.1}$$

For now we assume that the tuning functions can be described by unimodal functions of arbitrary shape, and generalize our analysis to the case of monotonic tuning curves of arbitrary shape in section 2.5.1.

The response model assumes that variance of the neural responses is directly proportional to the mean responses, which has been observed experimentally in some cases [85], but may not be true in general. The assumption that neuronal responses are statistically independent conditioned on the stimulus value is usually not correct [86, 87]. In section 2.5.2, we generalize our results to consider Poisson-like response models that belong to the exponential family of probability distributions with linear sufficient statistics [62, 88]. These distributions allow for stimulus dependent correlations and an arbitrary linear relationship between mean and variance of the population response.

We assume the total expected spike rate, R, of the population is fixed, which places a constraint on the tuning curves:

$$\int p(s) \sum_{n=1}^{N} h_n(s) \ \mathrm{d}s = R, \tag{2.2}$$

where $p(s)$ is the probability distribution of stimuli in the environment, and can have an arbitrary form. We refer to this as a sensory prior, in anticipation of its future use in Bayesian decoding of the population response.

2.3 Objective function

We now ask: what is the best way to represent values drawn from $p(s)$ given the limited resources of N neurons and R total spikes? To formulate a family of objective functions which depend on both $p(s)$, and the tuning curves, we first rely on Fisher information, $I_f(s)$, which is defined as [89]

$$I_f(s) = -\sum_{\vec{r}} p(\vec{r}|s)\ \frac{\partial^2}{\partial s^2} \log p(\vec{r}|s).$$

The Fisher information provides a measure of how accurately a population response represents a stimulus parameter based on the encoding model. It has been used to answer theoretical questions about the influence of tuning curve shapes [78, 79, 90] and response variability [91, 92] on the representational accuracy of population codes. It has also been used in neurophysiological studies to quantify changes in coding accuracy resulting from changes in tuning curve shapes during adaptation [93–95]. For the independent Poisson noise model, the Fisher information can be expressed analytically as [54]

$$I_f(s) = \sum_{n=1}^{N} \frac{h_n'^2(s)}{h_n(s)},$$

where $h_n'(s)$ is the derivative of the n^{th} tuning curve.

The Fisher information can also be used to express lower bounds on mutual information [83], the variance of an unbiased estimator [89], and perceptual discriminability [96]. Specifically, the mutual information, $I(\vec{r}; s)$, is bounded by:

$$I(\vec{r}; s) \geq H(s) - \frac{1}{2} \int p(s)\ \log\left(\frac{2\pi e}{I_f(s)}\right)\ \mathrm{d}s, \tag{2.3}$$

where $H(s)$ is the entropy, or amount of information inherent in $p(s)$, which is independent of the neural population. The bound is tight in the limit of low noise which can occur as N increases, R increases, or both [83].

The Cramer-Rao inequality allows us to express the minimum expected squared stimulus discriminability achievable by any decoder:

$$\delta^2 \geq \Delta^2 \int \frac{p(s)}{I_f(s)} \, \mathrm{d}s. \tag{2.4}$$

The constant Δ determines the performance level at threshold in a discrimination task. The conventional Cramer-Rao bound expresses the minimum mean squared error of any estimator, and in general requires a correction for the estimator bias [89]. Here, we use it to bound the squared *discriminability* of the estimator, as expressed in the stimulus space, which is independent of bias [96].

We formulate a generalized objective function that includes the Fisher bounds on information and discriminability as special cases:

$$\underset{h_n(s)}{\arg\max} \int p(s) \; f\left(\sum_{n=1}^{N} \frac{h_n'^2(s)}{h_n(s)}\right) \mathrm{d}s, \qquad \text{s.t.} \quad \int p(s) \sum_{n=1}^{N} h_n(s) \;\, \mathrm{d}s = R, \tag{2.5}$$

where $f(\cdot)$ is either the logarithm, or a power function. When $f(x) = \log(x)$, optimizing Eq. (2.5) is equivalent to maximizing the lower bound on mutual information given in Eq. (2.3). We refer to this as the *infomax* objective function. Otherwise, we assume $f(x) = x^\alpha$, for some exponent α. Optimizing Eq. (2.5) with $\alpha = -1$ is equivalent to minimizing the squared discriminability bound expressed in Eq. (2.4). We refer to this as the *discrimax* objective function.

2.4 How to optimize?

The objective function expressed in Eq. (2.5) is difficult to optimize because it is non-convex. To facilitate the optimization, we first parameterize a heterogeneous neural population by warping and rescaling a homogeneous population, as specified by a *cell density* function, $d(s)$, and a *gain* function, $g(s)$, that results in tuning widths that are inversely proportional to the cell density. Second, we show that Fisher information can be closely approximated as a continuous function of density and gain. Finally, re-writing the objective function and constraints in these terms allows us to obtain closed-form solutions for the optimal tuning curves.

2.4.1 Density and gain for a homogeneous population

If $p(s)$ is uniform, then by symmetry, the Fisher information for an optimal neural population should also be uniform. We assume a convolutional population of unimodal tuning curves, evenly spaced on the unit lattice, such that they approximately "tile" the space:

$$\sum_{n=1}^{N} h(s-n) \approx 1.$$

We also assume that this population has an approximately constant Fisher information:

$$\begin{aligned} I_f(s) &= \sum_{n=1}^{N} \frac{h'^2(s-n)}{h(s-n)} \\ &= \sum_{n=1}^{N} \phi(s-n) \approx I_{\text{conv}}. \end{aligned} \tag{2.6}$$

That is, we assume that the Fisher information curves for the individual neurons, $\phi(s-n)$, also tile the stimulus space. The value of the constant, I_{conv}, is dependent on the details of the tuning curve shape, $h(s)$, which we leave unspecified. As an example, Fig. 2.1(a-b) shows that the Fisher information for a convolutional population of Gaussian tuning curves, with appropriate width, is approximately

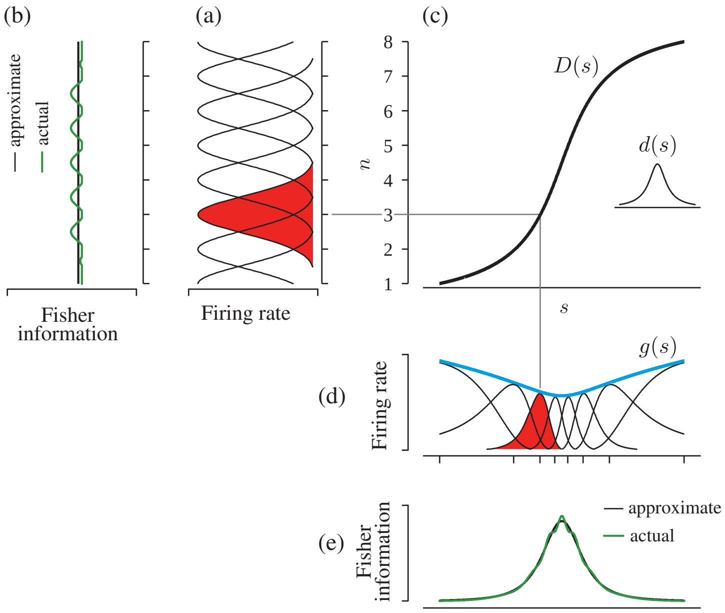

Figure 2.1: Construction of a heterogeneous population of neurons. **(a)** Homogeneous population with Gaussian tuning curves on the unit lattice. The tuning width of $\sigma = 0.55$ is chosen so that the curves approximately tile the stimulus space. **(b)** The Fisher information of the convolutional population (green) is approximately constant. **(c)** Inset shows $d(s)$, the tuning curve density. The cumulative integral of this density, $D(s)$, alters the positions and widths of the tuning curves in the convolutional population. **(d)** The warped population, with tuning curve peaks (aligned with tick marks, at locations $s_n = D^{-1}(n)$), is scaled by the gain function, $g(s)$ (blue). A single tuning curve is highlighted (red) to illustrate the effect of the warping and scaling operations. **(e)** The Fisher information of the inhomogeneous population is approximately proportional to $d^2(s)g(s)$.

constant. Now we introduce two scalar values, a gain (g), and a density (d), that affect the convolutional population as follows:

$$h_n(s) = g\ h\left(d(s - \frac{n}{d})\right). \tag{2.7}$$

The gain modulates the maximum average firing rate of each neuron in the population. The density controls both the spacing and width of the tuning curves: as the density increases, the tuning curves become narrower, and are spaced closer together so as to maintain their tiling of stimulus space. The effect of these two parameters on Fisher information is:

$$\begin{aligned} I_f(s) &= d^2 g \sum_{n=1}^{N(d)} \phi(ds - n) \\ &\approx d^2 g \; I_{\text{conv}}. \end{aligned}$$

The second line follows from the assumption of Eq. (2.6), that the Fisher information of the convolutional population is approximately constant with respect to s.

The total resources, N and R, naturally constrain d and g, respectively. If the original (unit-spacing) convolutional population is supported on the interval $(0, Q)$ of the stimulus space, then the number of neurons in the modulated population must be $N(d) = Qd$ to cover the same interval. Under the assumption that the tuning curves tile the stimulus space, Eq. (2.2) implies that $R = g$ for the modulated population.

2.4.2 Density and gain for a heterogeneous population

Intuitively, if $p(s)$ is non-uniform, the optimal Fisher information should also be non-uniform. This can be achieved through inhomogeneities in either the tuning curve density or gain. We thus generalize density and gain to be continuous functions of the stimulus, $d(s)$ and $g(s)$, that warp and scale the convolutional population:

$$h_n(s) = g(s_n)\ h(D(s) - n). \tag{2.8}$$

Here, $D(s) = \int_{-\infty}^{s} d(t)dt$, the cumulative integral of $d(s)$, warps the shape of the prototype tuning curve. The value $s_n = D^{-1}(n)$ represents the preferred stimulus value of the (warped) nth tuning curve (Fig. 2.1(b-d)). Note that the warped population retains the tiling properties of the original convolutional population. As in the uniform case, the density controls both the spacing and width of the tuning curves. This can be seen by rewriting Eq. (2.8) as a first-order Taylor expansion of $D(s)$ around s_n:

$$h_n(s) \approx g(s_n)\ h(d(s_n)(s - s_n)),$$

which is a generalization of Eq. (2.7).

We can now write the Fisher information of the heterogeneous population of neurons in Eq. (2.8) as

$$\begin{aligned} I_f(s) &= \sum_{n=1}^{N} d^2(s)\ g(s_n)\ \phi(D(s) - n) && (2.9) \\ &\approx d^2(s)\ g(s)\ I_{\text{conv}}. && (2.10) \end{aligned}$$

In addition to assuming that the Fisher information is approximately constant (Eq. (2.6)), we have also assumed that $g(s)$ is smooth relative to the width of $\phi(D(s)-n)$ for all n, so that we can approximate $g(s_n)$ as $g(s)$ and remove it from the sum. The end result is an approximation of Fisher information in terms of the continuous parameterization of cell density and gain. As earlier, the constant I_{conv} is determined by the precise shape of the tuning curves.

As in the homogeneous case, the global resource values N and R will place constraints on $d(s)$ and $g(s)$, respectively. In particular, we require that $D(\cdot)$ map the entire input space onto the range $[1, N]$, and thus $D(\infty) = N$, or equivalently, $\int d(s)\, \mathrm{d}s = N$. To attain the proper rate, we use the fact that the warped tuning curves sum to unity (before multiplication by the gain function) and use Eq. (2.2) to obtain the constraint $\int p(s) g(s)\, \mathrm{d}s = R$.

2.4.3 Objective function and solution for a heterogeneous population

Approximating Fisher information as proportional to squared density and gain allows us to re-write the objective function and resource constraints of Eq. (2.5) as

$$\underset{d(s),g(s)}{\arg\max} \int p(s)\ f\left(d^2(s)\ g(s)\right)\ \mathrm{d}s, \qquad \text{s.t.} \quad \int d(s)\, \mathrm{d}s = N, \tag{2.11}$$

$$\text{and} \quad \int p(s) g(s)\, \mathrm{d}s = R. \tag{2.12}$$

A closed-form optimum of this objective function is easily determined using calculus of variations. Specifically, one can compute the gradient of the Lagrangian, set to zero, and solve the resulting system of equations. Solutions are provided in Table 2.1 for the infomax, discrimax, and general power cases.

In all cases, the solution specifies a power-law relationship between the prior, and the density and gain of the tuning curves. In general, all solutions allocate more

		Infomax	**Discrimax**	**General**
Optimized function:		$f(x) = \log x$	$f(x) = -x^{-1}$	$f(x) = -x^{\alpha}$, $\alpha < \frac{1}{3}$
Density (Tuning width)$^{-1}$	$d(s)$	$Np(s)$	$\propto Np^{\frac{1}{2}}(s)$	$\propto Np^{\frac{\alpha-1}{3\alpha-1}}(s)$
Gain	$g(s)$	R	$\propto Rp^{-\frac{1}{2}}(s)$	$\propto Rp^{\frac{2\alpha}{1-3\alpha}}(s)$
Fisher information	$I_f(s)$	$\propto RN^2p^2(s)$	$\propto RN^2p^{\frac{1}{2}}(s)$	$\propto RN^2p^{\frac{2}{1-3\alpha}}(s)$
Discriminability bound	$\delta_{\min}(s)$	$\propto p^{-1}(s)$	$\propto p^{-\frac{1}{4}}(s)$	$\propto p^{\frac{1}{3\alpha-1}}(s)$

Table 2.1: Optimal heterogeneous population properties, for objective functions specified by Eq. (2.12).

neurons, with correspondingly narrower tuning curves, to higher-probability stimuli. In particular, the infomax solution allocates an approximately equal amount of probability mass to each neuron. The shape of the optimal gain function depends on the objective function: for $\alpha < 0$, neurons with lower firing rates are used to represent stimuli with higher probabilities, and for $\alpha > 0$, neurons with higher firing rates are used for stimuli with higher probabilities. Note also that the global resource values, N and R, enter only as scale factors on the overall solution, allowing us to easily test the validity of the predicted relationships on experimental data. In addition to power-law relationships between tuning properties and sensory priors, our formulation offers a direct relationship between the sensory prior and perceptual discriminability. This can be obtained by substituting the optimal solutions for $d(s)$ and $g(s)$ into Eq. (2.9), and using the resulting Fisher information to bound the discriminability, $\delta(s) \geq \delta_{\min}(s) \equiv \Delta/\sqrt{I_f(s)}$ [96]. The resulting expressions are provided in Table 2.1. In general, the solutions predict that discrimination thresholds should be lower for more frequently occurring stimuli.

2.5 Extensions

2.5.1 Monotonic tuning curves

Thus far we have solved for the optimal cell density and gain for warping and scaling a homogeneous population of unimodal tuning curves. However, many neurons exhibit *monotonic* tuning to intensity variables such as contrast, or sound pressure level. The influence of continuous cell density and gain on the Fisher information of a homogeneous population of monotonic tuning curves is the same as in the unimodal case (Eq. (2.10)), again assuming that the Fisher information curves of the homogeneous population tile. The constraint on N is also same. However, the total spiking cost fundamentally differs. Neurons with monotonic tuning curves saturate, and thus the entire population will be active at the high range of stimulus values, which incurs a large metabolic cost for encoding these values. Intuitively, this metabolic penalty can be reduced by lowering the gains of neurons tuned to the low end of the stimulus range, or by adjusting the cell density such that there are more tuning curves tuned to the high end of the stimulus range. It is unclear how the reductions in metabolic cost for these coding strategies may trade off with the optimal coding of sensory information.

To derive the optimal monotonic coding scheme, we first parameterize a heterogeneous population of monotonic tuning curves by warping and scaling the *derivatives* of a homogeneous population of monotonic tuning curves:

$$h_n(s) = \int_{-\infty}^{s} h'_n(t)\,\mathrm{d}t = \int_{-\infty}^{s} g(s_n)d(t)h'(D(t)-n)\,\mathrm{d}t. \tag{2.13}$$

This expression is similar to the parameterization of a heterogeneous population of unimodal tuning curves (Eq. (2.8)), except here, $h(\cdot)$ is now a prototype monotonic tuning curve. The density controls both the number of tuning curves and their slopes, which are inversely proportional to the cell density. The derivatives of the (warped) monotonic tuning curves, $h'(D(t)-n)$, will be unimodal functions, allowing us to use similar approximations and intuitions developed for the unimodal case. In particular, we assume that the derivatives of the tuning curves tile such that $\sum_{n=1}^{N} h'(D(t)-n) \approx 1$.

The total spike count can be expressed from Eqs. (2.2 & 2.13) as,

$$R = \int_{-\infty}^{\infty} p(s) \int_{-\infty}^{s} d(t) \sum_{n=1}^{N} g(s_n) h'(D(t)-n) \, \mathrm{d}t \, \mathrm{d}s.$$

We define a continuous version of the gain as $g(t) = \sum_{n=1}^{N} g(s_n) h'(D(t)-n)$ which allows us to approximate the total number of spikes as

$$\begin{aligned} R &= \int_{-\infty}^{\infty} p(s) \int_{-\infty}^{s} d(t) g(t) \\ &= \int_{-\infty}^{\infty} (1-P(s)) \, d(s) g(s) \, \mathrm{d}s \end{aligned}$$

In the second step, we performed integration by parts and defined $P(s) = \int_{-\infty}^{s} p(t) \, \mathrm{d}t$ as the cumulative density function of the sensory prior. The constraint on the total number of spikes is very different than the bell-shaped tuning curve case, as it now depends on the cell density and the cumulative distribution of the sensory prior, and will thus affect the optimal solutions for cell density and gain.

We reformulate the original optimization problem of Eq. (2.5) for monotonic

		Infomax	**Discrimax**	**General**
	Optimized:	$f(x) = \log x$	$f(x) = -x^{-1}$	$f(x) = -x^{\alpha}$, $\alpha < \frac{1}{3}$
Density	$d(s)$	$Np(s)$	$\propto Np(s)^{\frac{1}{3}} \left[1 - P(s)\right]^{\frac{1}{3}}$	$\propto Np(s)^{\frac{1}{1-2\alpha}} \left[1 - P(s)\right]^{\frac{\alpha}{2\alpha-1}}$
Gain	$g(s)$	$RN^{-1} \left[1 - P(s)\right]^{-1}$	$RN^{-1} \left[1 - P(s)\right]^{-1}$	$RN^{-1} \left[1 - P(s)\right]^{-1}$
Fisher.	$I_f(s)$	$\propto RNp^2(s) \left[1 - P(s)\right]^{-1}$	$\propto RNp^{\frac{2}{3}}(s) \left[1 - P(s)\right]^{-\frac{1}{3}}$	$\propto RNp^{\frac{2}{1-2\alpha}}(s) \left[1 - P(s)\right]^{\frac{1}{2\alpha-1}}$
Discrim.	$\delta_{\min}(s)$	$\propto p^{-1}(s) \left[1 - P(s)\right]^{\frac{1}{2}}$	$\propto p^{-\frac{1}{3}}(s) \left[1 - P(s)\right]^{\frac{1}{6}}$	$\propto p^{\frac{1}{2\alpha-1}}(s) \left[1 - P(s)\right]^{\frac{1}{2-4\alpha}}$

Table 2.2: Optimal heterogeneous population properties, for objective functions specified by Eq. (2.14).

tuning curves as:

$$\underset{d(s),g(s)}{\arg\max} \int p(s)\ f\left(d^2(s)\ g(s)\right)\ \mathrm{d}s, \qquad \text{s.t.} \quad \int d(s)\,\mathrm{d}s = N, \tag{2.14}$$
$$\text{and} \quad \int (1 - P(s))\, d(s) g(s)\,\mathrm{d}s = R.$$

A closed-form optimum of this objective function is easily determined by taking the gradient of the Lagrangian, setting to zero, and solving the resulting system of equations. Solutions are provided in Table. 2.2 for the infomax, discrimax, and general power cases, in addition to solutions for the optimal Fisher information and minimum achievable discrimination thresholds achievable by a subsequent perceptual system.

For all objective functions, the solutions for the optimal density, gain, and discriminability are products of power law functions of the sensory prior, and its cumulative distribution. In general, all solutions allocate more neurons with greater dynamic range to more frequently occurring stimuli. The optimal gain is the same in all cases. For a neuron tuned to a particular stimulus value, the optimal gain will be inversely proportional to the probability of all stimuli occurring after that stimulus value. Intuitively, this solution allocates lower gains to neurons tuned to the low end of the stimulus range, which is metabolically less costly. The global resource values again only appear as scale factors in the overall solution, allowing us to easily test the validity of the predicted relationships on experimental data.

2.5.2 Generalization to Poisson-like noise distributions

Our results depend on the assumption that neuronal variability is Poisson distributed and neural responses are statistically independent. To generalize our results, we also consider a model of neuronal variability that is "Poisson-like" and can include correlated neuronal variability [62, 88]. The probability density of the population response can be written as

$$p(\vec{r}|s) = f(\vec{r}) \exp\left[\boldsymbol{\eta}(s)^T \vec{r} - A(\boldsymbol{\eta})\right]. \tag{2.15}$$

This distribution belongs to the exponential family with linear sufficient statistics where the parameter $\boldsymbol{\eta}(s)$ is a matrix of the natural parameters of the distribution with the n^{th} column equal to $\eta_n(s)$, $A(\boldsymbol{\eta})$ is a (log) normalizing constant that ensures the distribution integrates to one, and $f(\vec{r})$ is an arbitrary function of the firing rates. The independent Poisson noise model considered in Eq. (2.1) is a member of this family of distributions with parameters: $\boldsymbol{\eta}(s) = \log \mathbf{h}(\mathbf{s})$ where $\mathbf{h}(\mathbf{s})$ is a matrix of tuning curves with the n^{th} column given $h_n(s)$, $A(\boldsymbol{\eta}) = \sum_{n=1}^{N} \exp(\eta_n)$, and $f(\vec{r}) = \prod_{n=1}^{N} \frac{1}{r_n!}$.

All of our objective functions depend on an analytical form for the Fisher information in terms of tuning curves, which is then expressed in terms of density and gain. To derive the Fisher information for the response model in Eq. (2.15), we start by noting that the derivative of natural parameters is related to the stimulus dependent covariance matrix of the population responses, $\Sigma(s)$, and the derivative of the tuning curves as [62, 88],

$$\frac{\partial \boldsymbol{\eta}}{\partial s} = \Sigma^{-1}(s)\frac{\partial \mathbf{h}}{\partial s}. \tag{2.16}$$

The term $\Sigma^{-1}(s)$ is the inverse of the covariance matrix, and is often referred to as a precision matrix.

The Fisher information *matrix* about the natural parameters is simply equal to the covariance matrix [89],

$$I_f\left[\boldsymbol{\eta}(s)\right] = \Sigma(s). \tag{2.17}$$

The local Fisher information about the stimulus, s, can be derived from the chain rule as,

$$I_f(s) = \frac{\partial \boldsymbol{\eta}}{\partial s}^T I_f\left[\boldsymbol{\eta}(s)\right]\frac{\partial \boldsymbol{\eta}}{\partial s}.$$

After substituting the relationships in Eq. (2.16 & 2.17) into this expression we obtain the final expression for the local Fisher information

$$I_f(s) = \frac{\partial \mathbf{h}}{\partial s}^T \Sigma^{-1}(s)\frac{\partial \mathbf{h}}{\partial s}. \tag{2.18}$$

The influence of Fisher information on coding accuracy is now directly dependent on knowledge of stimulus dependent (inverse) covariance matrix. Estimating such a precision matrix from experimental data is technically challenging (although see [87]). Here, we assume a biologically plausible precision matrix that allows for neuronal variability to be proportional to the mean firing rate, and the responses of nearby neurons to be correlated [91]. For a homogeneous neural population, $h_n(s) = h(s-n)$, we express each element in the precision matrix as,

$$\Sigma^{-1}_{n,m}(s) = \frac{\alpha\delta_{n,m} + \beta(\delta_{n,m+1} + \delta_{n+1,m})}{\sqrt{h(s-n)h(s-m)}}. \tag{2.19}$$

The parameter α controls a linear relationship between the mean response and the variance of the response for all the neurons. The parameter β controls the degree of the correlations, and $\delta_{n,n} = 1$ for all n while $\delta_{n,m} = 0$ if $n \neq m$. The Fisher information of a homogeneous population may now be expressed from Eqs. (2.18 &2.19) as,

$$\begin{aligned} I_f(s) &= \alpha \sum_{n=1}^{N} \phi(s-n) \quad + \beta \sum_{n,m=n\pm1} \frac{h'(s-n)h'(s-m)}{\sqrt{h(s-n)h(s-m)}} \\ &\approx \alpha I_{\text{conv}} + \beta I_{\text{corr}} \end{aligned}$$

In the last step we make two assumptions. First, we assume (as for the independent Poisson case) the Fisher information curves, $\phi(s-n)$, of the homogeneous population tile such that they sum to the constant, I_{conv}. Second, we assume that the cross terms, $\frac{h'(s-n)h'(s-m)}{\sqrt{h(s-n)h(s-m)}}$, also tile such that they sum to the constant, I_{corr}.

The Fisher information for a heterogeneous population, obtained by warping and scaling the homogeneous population by the density and gain is

$$\begin{aligned} I_f(s) &= d^2(s)\alpha \sum_{n=1}^{N} g(s_n)\phi(D(s)-n) \qquad (2.20) \\ &\quad + d^2(s)\beta \sum_{n,m=n\pm 1} \frac{g(s_n)g(s_m)}{\sqrt{g(s_n)g(s_m)}} \frac{h'(D(s)-n)h'(D(s)-m)}{\sqrt{h(D(s)-n)h(D(s)-m)}} \\ &\approx d^2(s)g(s)\left[\alpha I_{\text{conv}} + \beta I_{\text{corr}}\right]. \qquad (2.21) \end{aligned}$$

In the second step we make three assumptions. First, (as for the independent Poisson case) we assume $g(s)$ is smooth relative to the width of $\phi(D(s)-n)$ for all n, so that we can approximate $g(s_n)$ as $g(s)$. Second, we assume that the neurons are sufficiently dense such that $\frac{g(s_n)g(s_m)}{\sqrt{g(s_n)g(s_m)}} \approx g(s_n)$. Finally, we assume $g(s)$ is also smooth relative to the width of the cross terms. As a result, the gain factors can be approximated by same the continuous gain function, $g(s)$, and can be pulled out of both sums.

Given the form of the Fisher information (Eq. (2.21)), we conclude that the optimal solutions for the density and gain are the same as those expressed in Tables 2.2 & 2.1, which were derived for an independent Poisson noise model ($\alpha = 1, \beta = 0$). The absolute values of the Fisher information, and minimum achievable discrimination thresholds now depend on three additional scale factors, α, β, and I_{corr}, that characterize the correlated variability of the population code.

2.6 Discussion

We have examined the influence sensory priors on the optimal allocation of neural resources, as well as the influence of these optimized resources on subsequent perception. For a family of objective functions, we obtain closed-form solutions specifying power law relationships between the prior probability distribution of a sensory variable encountered in the environment, the tuning properties of a population that encodes that variable, and the minimum perceptual discrimination thresholds achievable for that variable. The solutions are easily testable with experimental data. In the next chapter, we show that the infomax predictions, for populations of neurons characterized by unimodal tuning curves, are remarkably consistent with environmental, physiological, and perceptual data.

Our analysis requires several approximations and assumptions in order to arrive at an analytical solution. We first rely on lower bounds on mutual information and discriminability based on Fisher information. Fisher information is known to provide a poor bound on mutual information when there are a small number of neurons, a short decoding time, or non-smooth tuning curves [83, 97]. It also provides a poor bound on supra-threshold discriminability [90, 98]. However, we do not require the bounds on either information or discriminability to be tight, but rather that their optima be close to that of their corresponding true objective functions. We also made several assumptions in deriving our results: (1) the tuning curves, $h(D(s) - n)$, or in the monotonic case their derivatives, $h'(D(s) - n)$, evenly tile the stimulus space; (2) the single neuron Fisher informations, $\phi(D(s) - n)$, evenly tile the stimulus space; and (3) the gain function, $g(s)$, varies slowly and smoothly over the width of $\phi(D(s) - n)$. These assumptions allow us to approximate Fisher information in terms of cell density and gain (Fig. 2.1(e)), to express the resource constraints in simple form, and to obtain a closed-form solution to the optimization problem.

Our framework offers an important generalization of the population coding literature, allowing for non-uniformity of sensory priors, and corresponding heterogeneity in tuning and gain properties. Nevertheless, it suffers from many of the same simplifications found in previous literature. First, tuning curve models only specify neural responses to a single stimulus values. The model should be generalized to handle arbitrary combinations of scalar inputs. Second, the response model should be extended beyond tuning curves to a description that can handle multi-dimensional sensory inputs such as images or sounds. Each of these limitations offers an important opportunity for future work.

Chapter 3

Efficient Sensory Coding Predicts the Heterogeneities of Neural Populations and Perception

3.1 Introduction

Neurons in sensory systems are often characterized in terms of their selectivity or 'tuning' for particular stimulus variables (e.g., acoustic frequency, or visual orientation). And perceptual experiments commonly characterize the ability of an observer to discriminate stimuli that differ in terms of these same stimulus variables. In both cases, the observed representation of these variables is typically heterogeneous: neural tuning properties and perceptual discriminability are different for different values of the stimulus variable. Despite the ubiquity of this observation, no current theory explains why this should be the case, or how these heterogeneities might be related to each other.

In chapter 2, we developed the ecologically-motivated explanation that these variations arise because of heterogeneities in the probability of encountering different stimuli in the natural environment. Specifically, we developed a variant of the theory of efficient coding [1, 2], which posits that sensory systems are optimized to extract and represent information about the sensory world, while minimizing metabolic resources. We showed that the optimal solution provides simple and testable predictions regarding the relationship between the frequency of occurrence of stimulus values, the neural selectivity for those values, and perceptual discriminability of those values. Here, we test this relationship for three visual and two auditory attributes, and find that it is remarkably consistent with existing data.

3.2 Results

Consider a stimulus variable, s, that is to be encoded in the responses of a population of N neurons, limited to a total spike rate of R. As is common in sensory neuroscience, the response of each neuron is characterized by a "tuning curve" that represents the average spike rate as a function of the stimulus value. For simplicity, we assume that neuronal response variability is Poisson distributed with the rate parameter defined by the tuning curve [54], and that for any stimulus, the responses of the neurons in the population are uncorrelated. Our results can also be generalized to a class of Poisson-like distributions that include correlations [62, 88] without changing the form of the result (see methods).

The information about the variable s represented by the noisy population responses increases with both N, and R [54,59,78,79]. But suppose the environment is inhomogeneous, in that the frequency of occurrence of the stimulus variable, as expressed by a probability distribution, $p(s)$, varies significantly over the range of s. Intuitively, a "good" sensory system would allocate a higher proportion of neurons or spikes (or both) to the most frequently occurring stimuli, improving the encoding accuracy of those stimuli at the cost of decreasing the accuracy of infrequently occurring stimuli. Formally, we seek the set of tuning curves that maximize the information conveyed about stimuli drawn from the distribution $p(s)$, subject to the two resource constraints N and R. To facilitate the optimization, we parameterize a heterogeneous neural population by warping and rescaling a homogeneous population, as specified by a *cell density* function, $d(s)$, and a *gain* function, $g(s)$, that results in tuning widths that are inversely proportional to the cell density: $w(s) \propto \frac{1}{d(s)}$ (Fig. 2.1).

We optimized the parameters of the population for the transmission of stimulus information, expressed using a lower bound based on Fisher information [83]. The Fisher information is proportional to the product of the squared cell density and the gain, and the lower bound on information is the expectation of the log Fisher information under the stimulus distribution [99]. We verify that this bound provides a good approximation under the conditions explored here (see methods). The resulting optimization problem, constrained by the two resources, is:

$$\operatorname*{arg\,max}_{d(s),g(s)} \int p(s) \log p \left(d^2(s) g(s) \right) \, \mathrm{d}s, \qquad \text{subject to} \quad \int d(s) \, \mathrm{d}s = N,$$

$$\text{and} \quad \int p(s) g(s) \, \mathrm{d}s = R,$$

and a closed form solution is readily obtained using calculus of variations:

$$d(s) = Np(s), \qquad w(s) \propto \frac{1}{d(s)} = \frac{1}{Np(s)}, \qquad g(s) = R. \tag{3.1}$$

The structure of this optimal population directly reflects the statistical properties of the environment. Specifically, the cell density is proportional to the stimulus distribution, ensuring that frequently occurring stimuli are coded with greater precision, using denser and more narrowly tuned cells. On the other hand, we see that the maximal response (gain) of the cells in the optimal population is constant, independent of the preferred stimulus value. Since we have assumed the tuning widths are inversely proportional to cell density, and thus to the stimulus distribution, this solution implies that the average response of each neuron (over stimuli encountered in the world), is identical across the population. Finally, the unknown total resource values $\{N, R\}$ appear only as multiplicative scale factors in the expressions for gain and density, and thus the optimal solution provides a unique prediction for the shapes of both the cell density and tuning width as a function of preferred stimulus.

The optimal population also limits the best achievable discrimination performance of a perceptual system that bases its responses on the output of this population. Specifically, the Fisher information, expressed in terms of cell density and gain, provides a lower bound on discriminability [54], even when the observer is biased [96]. Substituting the optimal cell density and gain into this bound gives an expression for the minimum achievable discrimination thresholds:

$$\delta_{\min}(s) \propto \frac{1}{\sqrt{d^2(s)g(s)}} = \frac{1}{N\sqrt{R}p(s)} \tag{3.2}$$

Thus, our solution predicts that frequently occurring stimuli should be more discriminable (specifically, inverse discrimination thresholds should be proportional to the probability of encountering a stimulus value). The shape of this solution is again a simple function of the stimulus probability, $p(s)$, scaled by a multiplicative factor that depends on neural resources and an additional factor that depends on the experimental conditions under which discrimination thresholds are measured (e.g., criterion value, stimulus duration, or intensity). As a result, the solution provides a unique prediction of the shape of perceptual discrimination as a function of stimulus value.

Our efficient coding framework predicts explicit relationships between sensory statistics, physiological tuning properties, and perceptual discriminability. We tested these relationships in the context of two auditory attributes (acoustic frequency and modulation frequency), and three visual attributes (local orientation, spatial frequency, and retinal speed). The data, and predictions, are shown in Fig. 3.1. Each row of this figure corresponds to data obtained for a particular attribute. Each of these attributes exhibit substantial heterogeneity in their statistical, physiological, and perceptual representations. Data in the first column (Fig. 3.1a-e) correspond to stimulus distributions for each attribute, as estimated from large databases of photographic images or sounds obtained from natural environments. Physiological data (Fig. 3.1f-j) are taken from single-cell electrophysiological recordings in primate or cat that report the independently measured tuning widths of a large population of neurons as a function of their preferred stimuli. These measurements were made in a diverse set of brain areas (the auditory nerve fibers, the inferior colliculus, the primary visual cortex, and the middle temporal cortex), each chosen based on a substantial literature identifying the tuning properties of those neurons for the stimulus feature of interest. For the case of local orientation, we also analyzed another physiological data set in which tuning widths are reported (see supplementary information). Estimates of the cell density in each area (Fig. 3.1k-o) are obtained with a histogram binned over the preferred stimuli. Discrimination thresholds for each sensory attribute (Fig. 3.1p-t) were measured in human perceptual experiments. In some cases, we include two perceptual data sets (distinguished by color) obtained under different experimental conditions.

For each attribute, we find that the predicted relationships between the environment, physiology, and perception (Fig. 3.1 thick black lines) are consistent with the data. In most cases, we used the histogram of the environmental data as an estimate of $p(s)$, and then used this to predict the physiological and perceptual data. Predicted curves for tuning width and perceptual discriminability are individually re-scaled to best match the corresponding data (since the true scale factors depend on the unknown values of N and R). In the case of local image speed (Fig. 3.1e,j,o,&t) for which environmental data are technically difficult to estimate, we used the theory in reverse, fitting the perceptual data and using this to generate a prediction for $p(s)$. For all other sensory attributes, we find that predictions of perceptual discrimination data are remarkably accurate. For all attributes, both physiological predictions are supported by the data and consistent with our assumption that tuning widths are inversely proportional to cell density. For tuning widths, the predictions account for $(39.6, 69.1, 47.7, 67.4)\%$ of the variance in the data, which corresponds to $(95.8, 93.5, 97.1, 98.8)\%$ of the data variance that is accounted for by the best-fitting power law with two additional free parameters. The predicted cell densities exhibit systematic deviations from the estimates near the two ends of the stimulus range (see supplementary information). Estimates of cell density may not be properly represented in the data, as they are limited by sample size and potential biases in electrode sampling. Finally, we examined the gain of several neural populations (see supplementary information). We find that, although there is significant variation in these values across the population, it is not systematically related to the stimulus, in accordance with the predictions of our framework.

The neural data exhibit scatter about their predicted values, and we wondered how much of an effect this variability would have on the information that they transmit. To quantify this, we compared the amount of information transmitted by each observed neural population, to that transmitted by the theoretically optimal populations (see methods). We find that, relative to the information transmitted by a homogeneous population with the same resources, the observed neural populations encode between $85 - 95\%$ of the information that would be transmitted by the optimal population (Fig. 3.2). Thus, despite the variability of the physiological data, we conclude that neural populations are near-optimal in their efficiency for transmitting signals drawn from their associated environmental distributions.

3.3 Discussion

The notion that an organism is adapted both neurophysiologically and perceptually to the statistics of the natural environment is of fundamental importance to evolutionary biology. Our framework instantiates this notion in a mathematically precise form, leading to a direct relationship between environmental distributions, the tuning properties of neural populations, and perceptual discriminability. We find that these predictions are supported by physiological data from a diverse set of brain areas, as well as human perceptual data.

These results generalize and extend a number of previously published results on specific variants of the efficient coding hypothesis. For populations of identical bell-shaped tuning curves, a non-uniform distribution of preferred stimuli was found to be optimal [83, 84]. However, these results were not compared directly to physiological data, and because they assumed uniform tuning widths, cannot account for perceptual discriminability (Fig. 3.1p-t). A number of studies have examined coding efficiency for single neurons with *monotonic* response functions, demonstrating that the optimal solution is proportional to the cumulative stimulus distribution [6–8, 11], consistent with some physiological data [6]. Our framework should facilitiate an extension of our results to populations of monotonic tuning curves. The optimal cell density and gain will surely differ from those presented here, since the constraint on total spike rate will depend on the cell density.

As with many previous studies of neural population coding, our results are limited to the description of single stimulus variables, and generalization to the joint encoding of multiple attributes is not straightforward. Nevertheless, we find that our physiological predictions are consistent with populations of linear receptive fields are that are numerically optimized to encode ensembles of natural images [14, 18, 27] or sounds [28]. Specifically, the tuning characteristics of these optimized receptive fields are consistent with our predictions of cell density and tuning width for the encoding of orientation, spatial frequency, and acoustic frequency (see supplementary information). We also derived and tested the predictions of an alternative objective function − that the neural systems may be optimized for discrimination performance. However, we find that the alternative predictions are not supported by the data sets examined here (see supplementary information).

Many have argued that coding efficiency is a reasonable objective for early stages of sensory processing, but seems unlikely to explain more specialized later stages responsible for producing actions [30]. Nevertheless, if we take seriously the notion that perception is a process of inference [4], then these later stages must rely on knowledge of the frequency of occurrence of sensory attributes in the environment. Although such prior information has been widely used in formulating Bayesian explanations for perceptual phenomena [32], the means by which it is represented within the brain is currently unknown [46]. The results presented here provide a potential solution, in which prior probabilities are implicitly embedded in the arrangement and selectivity of tuning curves. In Chapter 4 we show that the responses of an efficient population can be decoded in a biologically plausible mechanism that correctly combines current sensory information with the embedded prior. Thus, an efficient coding strategy may offer unforeseen benefits for explaining later stages of sensory processing.

3.4 Methods

3.4.1 Estimating environmental distributions

The environmental distributions for the two auditory attributes (acoustic frequency and modulation frequency) were computed from commercially available compilations of animal vocalizations (58 min) [100,101], background environmental sounds (113 min) [102], and recordings made while walking around a suburban university campus (62 min). The campus sounds were provided by Josh McDermott and were recorded with a Sennheiser omnidirectional microphone (ME62) and a Marantz solid state recorder (PMD670). The distributions for local orientation and spatial frequency were computed from three publicly available image databases comprised of a total of 816 natural scenes [15,103,104]. The distribution for speed was predicted according to the theory from the psychophysical data.

Acoustic Frequency

We assume that the ensemble power spectral density of sounds reflects the probability of acoustic frequencies occurring in the natural environment. We computed the power spectral density for each sound file in the database using Welch's method [118], with non-overlapping 500 millisecond segments windowed with a hamming filter to mitigate boundary effects. The ensemble power spectrum, $S(f)$, was fit to all recordings with a modified power-law [119]:

$$S(f) = \frac{A}{f_0^p + f^p}. \tag{3.3}$$

The parameters were chosen to minimize squared error to the data ($A = 2.4 \times 10^6, f_0 = 1.52 \times 10^3, p = 2.61$).

Modulation Frequency

We assume that the ensemble modulation power spectral density of sounds reflects the probability of modulation frequencies occurring in the natural environment. Each sound in the auditory database was decomposed into subbands using a physiologically motived bank of 30 raised cosine filters [120]. The center frequencies of the filters were equally spaced on an equivalent rectangular bandwidth ($\mathrm{ERB_N}$) scale, and the filter bandwidths (as a function of center frequency) were comparable to those of the human ear [121]. The temporal envelope of the output of each frequency channel was extracted by computing the magnitude of the analytic signal. The temporal modulation power spectrum was computed by averaging the power spectral density of each envelope across all frequency channels. The modulation spectrum of the envelope of a bandpass filter output is inevitably low-pass (with a cutoff determined by the filter bandwidth). To avoid biasing our measurements of modulation statistics, we only included frequencies below this filter cutoff in our average. The ensemble modulation spectrum was fit to all recordings with a modified power law, (Eq. 3.3), with parameters chosen to minimize squared error to the data ($A = 0.06, f_0 = 0, p = 0.84$).

Local Orientation

We used a Gaussian pyramid [122] to decompose each image in the database into a spatial scale ($2 - 5$ cycles/deg) that matched that of the grating stimuli used in the orientation discrimination experiment (4 cycles/deg) (Fig. 3.1r) [36].

The horizontal and vertical gradients, centered on each pixel in the resulting image, were computed with local rotation-invariant 5-tap derivative filters [123]. We computed an orientation tensor [124], defined as the covariance matrix of the gradients pooled across a local region. A pooling size of 1 degree was chosen to best match the physiological data, which was measured foveally (between $0-3$ deg retinal eccentricity) (Fig. 3.1m,h) [109]. We computed three quantities from the eigenvector decomposition of the orientation tensor: the energy (sum of the eigenvalues), orientedness (ratio of the eigenvalue difference to the eigenvalue sum), and the dominant orientation (angle of the eigenvector with the larger eigenvalue). We formed a histogram of the dominant orientations of all tensors for which the energy exceeded the 68th percentile of all energies in the database, and the orientedness exceeded 0.8. The histogram was converted to a probability distribution by normalizing by the total number of tensors that exceeded the two thresholds. We verified that the resulting distribution did not change significantly for modest changes in both thresholds.

Spatial Frequency

We assume that the radially integrated power spectral density of natural images reflects the probability of spatial frequencies occurring in the natural environment. The power spectral density for each image in the database was computed by taking the magnitude of the windowed Fast Fourier transform of the image, and integrating the result over orientation. The units of the power spectrum were converted from cycles per pixel to cycles per degree of visual angle by using the appropriate camera settings for which each image was captured. The ensemble spectra were fit with a modified power law (Eq. 3.3) with parameters chosen to minimize squared error to the data ($A = 0.21, f_0 = 0.11, p = 1.14$). The approximate $\frac{1}{f}$ nature of the estimated spectrum is consistent with many previous studies [e.g.,125, 126]).

Speed

The psychophysical data were fit with a power law,

$$\delta(s) = as^p + b,$$

with parameters chosen to minimize squared error to the data ($a = 0.05, p = 0.93, b = 0.11$). According to the theory (Eq. 3.2), the predicted environmental distribution for speed is estimated as $p(s) \propto \delta(s)^{-1}$ (Fig. 3.1e). This "slow speed prior" is qualitatively consistent with previous estimates of the distribution of retinal speed [35, 127, 128].

3.4.2 Cell density estimates

To quantitatively determine whether the estimated environmental distributions provide an accurate prediction of the physiologically measured cell densities, we performed a one sample Kolmogorov-Smirnov (KS) test. For our purposes, the KS statistic quantifies the maximal difference between the cumulative density function (CDF) of the environmental distribution, and the empirical CDF of the cell density (Fig. 3.3). For the attributes of acoustic frequency, modulation frequency, and spatial frequency, we found that the cell densities deviated significantly from the corresponding environmental distributions ($p < 0.001$ for each attribute, KS test). For speed, we cannot reject the null hypothesis that the cell density does not deviate from the estimated environmental distribution ($p = 0.15$, KS test). For all attributes, we find that the data are much closer to our predictions than a uniform distribution (Fig. 3.3 grey lines)

3.4.3 Estimating mutual information

Analytically computing the mutual information between the stimulus and the population response, $I(\vec{r}; s)$, is intractable. Instead, we estimate it with a Monte-Carlo approximation, assuming that neural responses are described as samples from an independent Poisson process with a rate parameter determined by each neurons tuning curve. To develop this procedure, we first express the mutual information as the difference between the stimulus entropy, $H(s)$, and the reduction in stimulus entropy once a population response is observed, $H(s|\vec{r})$:

$$I(\vec{r}; s) = H(s) - H(s|\vec{r})$$

The stimulus entropy is defined as,

$$H(s) = -\int p(s) \log p(s) \, \mathrm{d}s,$$

and, can be readily computed by numerical integration assuming $p(s)$ is known. The conditional entropy is defined as,

$$H(s|\vec{r}) = -\int\int p(\vec{r}, s) \log p(s|\vec{r}) \, \mathrm{d}s \, \mathrm{d}\vec{r}. \tag{3.4}$$

The conditional entropy is difficult to compute as it requires multi-dimensional integration. We instead approximate it with L samples from the joint distribution,

$$H(s|\vec{r}) \approx \frac{1}{L}\sum_{l=1}^{L} \log p(s_l|\vec{r}_l). \tag{3.5}$$

Here, $(s_l, \vec{r}_l) \sim p(s, \vec{r})$ denotes the l^{th} draw from the joint distribution. These samples can be readily obtained by an ancestral sampling procedure [129]. Ancestral sampling works by first drawing a sample from the prior distribution, $s_l \sim p(s)$, then drawing a sample from the conditional distribution, evaluated at this sample, $\vec{r}_l \sim p(\vec{r}|s_l)$. Since we have assumed that the response distribution is independent Poisson, the n^{th} element of $\vec{r}_l$ is a sample from a Poisson distribution with a rate parameter determined by its tuning curve evaluated at s_l: $r_{l,n} \sim \text{Poiss}(h_n(s_l))$. Each repeat of this procedure generates one independent sample from the joint distribution. Finally, we compute the posterior distribution required in Eq. (3.5) from Bayes rule as,

$$p(s_l|\vec{r}_l) = \frac{p(s_l)p(\vec{r}_l|s_l)}{\int p(s)p(\vec{r}_l|s)\,\mathrm{d}s}.$$

The numerator is straightforward to evaluate, and the denominator can be computed by numerical integration.

3.4.4 Estimating normalized information

The normalized mutual information, $I_{\text{norm}}(\vec{r}; s)$, is defined as:

$$I_{\text{norm}}(\vec{r}; s) = \frac{I_{\text{dat}}(\vec{r}; s) - I_{\text{hom}}(\vec{r}|s)}{I_{\text{het}}(\vec{r}; s) - I_{\text{hom}}(\vec{r}; s)}$$

Here, $I_{\text{hom}}(\vec{r}; s)$ is the information transmitted by the homogeneous population, $I_{\text{het}}(\vec{r}; s)$ is the information transmitted by the optimal heterogeneous population, and $I_{\text{dat}}(\vec{r}|s)$ is the information transmitted by the observed neural populations. A value of 0 indicates measured neural populations transmitting as much information as a homogeneous population. A value of 1 indicates that the measured neural population transmits as much information as the optimal heterogeneous population. The normalized information is not necessarily bounded between 0 and 1. Negative values indicate that the measured neural populations are transmitting less information than the homogenous populations, and values greater than 1 indicate that the measured neural populations transmit more information than the optimal solution.

To compute the normalized information for each attribute (Fig. 3.2) we first constructed a homogeneous population, characterized by Gaussian tuning curves (with the same width parameter) evenly spaced across the domain of the sensory prior. The number of tuning curves, N, was chosen to be equal to the number of cells observed for each attribute. The total number of spikes, R, was constrained to lie within a metabolically relevant regime [24]: $\frac{R}{N} = 0.1, 1,$ or 10 spikes per neuron. The width parameter of the Gaussian was chosen such that, after warping the homogeneous population by the optimal density function, the tuning widths of the resulting optimal heterogeneous population (measured as the full width at half maximum value) were equal to the predicted tuning widths (Fig. 3.1f-j thick black lines). For the physiological datasets, we did not have access to the empirically measured acoustic or modulation frequency tuning curves. We chose to model them with Gamma functions, with rate and shape parameters chosen to minimize squared error between the data and the widths and preferred stimuli of the Gamma functions. In the case of spatial frequency and speed, we used the measured tuning curves, which were fit to the data with a log Gaussian function [130]. These tuning curves exhibited substantial variability in their gains (Fig. 3.5). To ensure a comparison of information transmission for identical resources, we scaled the gains of tuning curves in each of these populations by a single value such that R was the same as in the corresponding homogeneous and optimal heterogeneous populations.

Once the tuning curves were specified, we estimated the normalized mutual information with the ancestral sampling procedure with $L = 10,000$ samples. We verified that this was a sufficient number of samples by repeating the procedure with $100,000$ samples and obtaining similar results. We also verified that the precise shape of the tuning curve used to model the physiological data for acoustic and modulation frequency did not significantly influence the normalized information by re-running the analysis with Gaussian and raised cosine tuning functions fit to the data. 95% confidence intervals for the normalized information were obtained with $1,000$ bootstrap estimates of $I_{\text{dat}}(\vec{r}|s)$. For all atributes and resource constraints (except speed with $\frac{R}{N} = 0.1$), the normalized information was significantly greater than 0 and very close to 1, indicating that the observed neural populations are close to optimal for information transmission despite significant heterogeneity in cell density, tuning width, and (for spatial frequency and speed) gain. Median values for the normalized information are presented in Table 3.1.

Spikes/Neuron (R/N):	0.1	1	10
Acoustic Frequency, $N = 553$	0.87	0.93	0.96
Modulation Frequency, $N = 262$	0.85	0.85	0.88
Spatial Frequency, $N = 538$	0.80	0.84	0.98
Speed, $N = 76$	0.26	0.97	1.03

Table 3.1: Normalized information, $I_{\text{norm}}(\vec{r}; s)$, computed for each data set under different resource constraints.

3.4.5 Testing the Fisher bound on mutual information

Our results are optimized for a lower bound on mutual information, based on Fisher information $I_f(s)$,

$$I(\vec{r}; s) \geq H(s) + \frac{1}{2} \int p(s) \log \left(\frac{I_f(s)}{2\pi e} \right) \mathrm{d}s. \tag{3.6}$$

where the Fisher information, assuming an independent Poisson noise model is give by [54].

$$I_f(s) = \sum_{n=1}^{N} \frac{h_n'^2(s)}{h_n(s)},$$

We verified that the lower bound on mutual information is indeed tight for the sensory priors, tuning curves, and resource constraints explored here. Therefore, our reliance on the lower bound for analytical tractability does not influence our optimal solutions. To see this, we computed the difference between the ancestral sampling estimate of mutual information for each data set, over a range of resource constraints, and the corresponding lower bound from Eq. 3.6. The difference between these values are shown in Table 3.2. In the worst case, the lower bound differed by a tenth of a bit. In all other cases, the bound underestimated the true

Spikes/Neuron (R/N):	0.1	1	10
Acoustic Frequency, $N = 553$	0.002	0.007	0.0001
Modulation Frequency, $N = 262$	0.07	0.02	0.008
Spatial Frequency, $N = 538$	0.03	0.006	0.001
Speed, $N = 76$	0.1	0.05	0.02

Table 3.2: Difference between mutual information and the lower bound (in bits) computed for each data set under different resource constraints.

information by less than a hundredth of a bit.

3.5 Supplementary information

3.5.1 Another orientation example

Visually oriented stimuli are more perceptually discriminable about horizontal or vertical axes rather than oblique axes. This "oblique" effect has been confirmed empirically in numerous behavioral studies in humans and animals published over the past century [131]. Our theoretical explanation for the oblique effect states that orientation discrimination thresholds should be inversely proportional to the frequency of occurrence of orientations in the natural environment. We found that this perceptual prediction is remarkably accurate (Fig. 3.1r). However, it relies on the underlying physiological prediction that there are more cells with narrower tuning for representing more frequently occurring orientations. Although we have shown that is partially true for one data set (Fig. 3.1m), other neurophysiological investigations into the anisotropy of orientation tuning preferences have provided mixed results with some groups finding strong anisotropies and others none (see [132] for review).

We analyzed a separate large data set of the orientation tuning properties of a population simple cells recorded from cat area 17 [132] (Fig. 3.4). There are indeed more cells with narrower tuning for more frequently occurring orientations. However, the magnitudes of these biases are significantly smaller than those found in the Macaque V1 data reported in Fig. 3.1 m and are inconsistent with the magnitude of the environmental and perceptual biases (Fig. 3.4 thick black lines).

There are several possible reasons for inconsistencies between the orientation data sets considered here. First, the measurements are made in different species, and may reflect differences in either the neurophysiology or visual environments of cats and monkeys. Second, the Macaque cells were recorded from the fovea. In the same study, it was found that the density of cells recorded in the periphery was uniform [109]. The cat data were recorded over a wide range of eccentricities, possibly diminishing the magnitude of the reported bias in cell density. Finally, there is theoretical and physiological evidence that orientation tuning preferences may vary spatially across the visual field [133, 134]. As a result orientation tuning preferences may also depend heavily on the angular position of the receptive fields relative to the fovea.

3.5.2 Predictions of gain

We examined the gain of V1 neurons tuned to spatial frequency [107], and MT neurons tuned to speed [108] (Fig. 3.5). We find that, although there is significant variation in these values for each attribute, it is not systematically related to the stimulus, in accordance with the predictions of our framework. This is also true of the orientation tuning data obtained from cat area 17 [132] (Fig. 3.4c). For both attributes we found a weak but insignificant linear correlation between preferred stimulus value and gain of each neuron ($r = 0.065, p = 0.13$ for spatial frequency and $r = 0.104, p = 0.37$ for speed).

To test for possible nonlinear relationships between the gain, g, and preferred stimuli, μ, we constructed null model that assumes statistical independence between the two quantities: $p(\mu, g) = p(\mu)p(g)$. The distribution for $p(\mu)$ was parameterized by an exponential distribution with mean parameter 1.53 cpd for spatial frequency and 17.35 deg/sec for speed, chosen to maximize the log likelihood of the data. The distribution for $p(g)$ was also parameterized by an exponential distribution with maximum likelihood parameters of 27.01 spikes/sec for spatial frequency and 29.71 spikes/sec for speed. For each attribute, we generated samples of preferred stimuli and gain from the null model, matched to the sample size of the data. We then computed a distribution of the log likelihood of the synthetic data under the null model by drawing new samples from the model and computing the likelihood $10,000$ times. For each attribute, we found that the log likelihood of the data under the independent model was well within the 95% confidence intervals computed from the distribution of log likelihoods of synthetic data sampled from the model. Therefore we cannot reject the hypothesis that the gain and preferred stimuli of the cells are statistically independent.

3.5.3 Predictions of alternative efficient coding frameworks

We find that our physiological predictions are qualitatively consistent with populations of linear receptive fields that are numerically optimized to encode ensembles of natural images [14, 27] or sounds [28]. Specifically, the tuning characteristics of these optimized receptive fields are consistent with our predictions of cell density and tuning width for the encoding of orientation, spatial frequency, and

	Input	**Response**	**Neural noise**	**Objective**	**Constraints**
This work	$p(s)$	tuning curves	Poisson	information	$\{N, R\}$
ICA[14,15]	natural images	linear RF	none	information	power
Sparse Coding [28]	natural sounds	linear RF	none	squared error	sparsity

Table 3.3: Comparison of efficient coding frameworks

acoustic frequency (Fig. 3.6). The visual receptive fields are derived from independent components analysis (ICA) and the auditory receptive fields are derived from sparse coding. These theories may be interpreted as variants of the efficient coding hypothesis, each using a different objective function and constraints, and making different assumptions about input distributions, neural response properties, and noise. These differences are summarized in Table 3.3, and further elaborated in the following sections.

Independent components analysis of natural image patches

ICA assumes a linear generative model of natural images,

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{3.7}$$

where $\mathbf{x}$ is a matrix with each column representing a different natural image, $\mathbf{A}$ is a matrix of basis functions, and $\mathbf{s}$ is a matrix of coefficients. The goal of ICA is to simultaneously learn the basis functions, $\mathbf{A}$, and coefficients, $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$, such that the coefficients are as statistically independent as possible. Assuming no neuronal noise, this procedure is equivalent to picking the neural responses, (often equated to the coefficients $\mathbf{s}$), and receptive fields (often equated to the columns of $\mathbf{A}$) to maximize the information transmitted about the natural image ensemble. To obtain a solution, we ran the Fast ICA algorithm [135] (with the default parameters) using one million $16X16$ pixel image patches sampled randomly from the same database of natural images used to compute the environmental distributions for local orientation and spatial frequency [15, 103]. The dimensionality of each patch was reduced by a factor of two using principle components analysis, effectively low-pass filtering the images, in order to aid the algorithms convergence. As described in previous literature, the optimal receptive fields (Fig. 3.6a) are seen to closely resemble those of simple cells in V1 [14, 15].

To derive the orientation and spatial frequency tuning properties of these receptive fields, we first computed the magnitude of the two dimensional Fourier transform of each filter and found the location of the maximum amplitude. We estimated the orientation tuning curve by interpolating the values of the magnitude as a function of angle about this peak value. A spatial frequency tuning curve was computed by interpolating the values of the magnitude radially through the location of the peak magnitude. We found that 98 of the 128 receptive fields exhibited clear tuning to both orientation and spatial frequency. For these units, we computed the bandwidths of the derived tuning curves as the full width at half maximum (Fig. 3.6b,d). The preferred stimuli of the tuning curves were computed as the orientation or spatial frequency that elicited the maximum response. A histogram of the preferred stimuli was used as an estimate for the local cell density (Fig. 3.6c,e). For both attributes, we find that the predictions of our framework are qualitatively consistent with the derived tuning characteristics (Fig. 3.6b-e thick black lines).

Sparse coding of natural sounds

Sparse coding also assumes a linear generative model of the input (Eq. 3.7). However, the goal of sparse coding is to learn the basis functions and coefficients to minimize reconstruction error subject to a sparsity constraint on the coefficients:

$$\underset{\mathbf{A},\mathbf{s}}{\arg\min} \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_2 + \lambda \|\mathbf{s}\|_0. \tag{3.8}$$

The first term in the objective function is the mean squared error between the generative model and the inputs. The second term enforces sparsity, or rather that there are as few active neurons as possible. The parameter λ controls the amount of tradeoff between sparsity and reconstruction error.

When the inputs consist of natural sounds, an approximate numerical solution to this objective function yields receptive fields that resemble those of auditory nerve fibers [28] (Fig. 3.6f). The tuning widths of these filters, as function of their preferred stimulus values, closely match the predictions of our efficient coding framework (Fig. 3.6g). We computed a histogram of the preferred stimuli as a local estimate of cell density and find that it is also qualitatively consistent with our results (Fig. 3.6 h).

3.5.4 Predictions of an alternative optimality principle

To further examine whether coding efficiency is a reasonable objective function for early sensory systems, we formulate and test an alternative hypothesis − that sensory systems are optimized to discriminate stimulus values [7]. Specifically, we consider the optimization of a lower bound on average squared stimulus discriminability [99], expressed in terms of cell density and gain. The new objective function and constraints are,

$$\operatorname*{arg\,max}_{d(s),g(s)} \int p(s) d^{-2}(s) g^{-1}(s)\, \mathrm{d}s, \qquad \text{subject to} \quad \int d(s)\, \mathrm{d}s = N,$$
$$\text{and} \quad \int p(s) g(s)\, \mathrm{d}s = R,$$

and a closed form solution is readily obtained using calculus of variations:

$$d(s) \propto N p^{\frac{1}{2}}(s), \qquad w(s) \propto \frac{1}{d(s)} = \frac{1}{N p^{\frac{1}{2}}(s)}, \qquad g(s) \propto \frac{R}{p^{\frac{1}{2}}(s)}. \tag{3.9}$$

The structure of this optimal population differs significantly from that of the optimally efficient population. Specifically, the cell density is proportional to the *square root* of stimulus distribution, thus allocating more cells, relative to the efficient coding solution, to less frequently occurring stimuli at the cost of allocating relatively less cells to more frequently occurring stimuli. The gain of the cells in the optimal population is inversely proportional to the square root of the stimulus distribution. Since we have assumed the tuning widths are inversely proportional to cell density, and thus to the square root of the stimulus distribution, this solution implies that the average response of each neuron (over stimuli encountered in the world), is identical across the population, as in the efficient coding solution. Finally, the unknown total resource values $\{N, R\}$ again appear only as multiplicative scale factors in the expressions for gain and density, and thus the optimal solution provides a unique prediction for the shapes of both the cell density and tuning width as a function of preferred stimulus.

Analogous to the efficient coding solution, this optimal population also limits the best achievable discrimination performance of a perceptual system that bases its responses on the output of this population. Specifically, the minimum achievable discrimination thresholds are expressed as,

$$\delta_{\min}(s) \propto \frac{1}{\sqrt{d^2(s)g(s)}} = \frac{1}{N\sqrt{R}p^{\frac{1}{4}}(s)}. \tag{3.10}$$

As in the efficient coding case, this solution predicts that frequently occurring stimuli should be more discriminable. However, this solution achieves lower average squared discrimination thresholds by producing lower thresholds, relative to the efficient coding solution, to less frequently occurring stimuli at the cost of producing relatively higher thresholds to more frequently occurring stimuli. The solution is again a simple function of the stimulus probability, $p(s)$, scaled by a multiplicative factor that depends on neural resources and an additional factor that depends on the experimental conditions under which discrimination thresholds are measured. As a result, the solution provides a unique prediction of the shape of perceptual discriminability as a function of stimulus value.

We compared these relationships to those predicted by efficient coding in the context of the previously discussed attributes. The data, and predictions, are shown in Fig. 3.7. For each attribute, we find that the predicted relationships between the environment, physiology, and perception for the optimal discriminability hypothesis (Fig. 3.7 thick green lines) are deviate more significantly from the data than the predictions of the efficient coding hypothesis (Fig. 3.7 thick black lines). For tuning widths, the new predictions account for $(32.0, 55.8, 36.0, 55.0)\%$ of the variance in the data, which corresponds to $(7.6, 13.32, 11.7, 12.4)\%$ less of the data variance that is accounted for by the efficient coding solutions. The predicted cell densities also deviate significantly from data ($p < 0.001$ for all attributes, one sample KS test). Finally, the new predictions of gain cannot account for the lack of systematic relationship between the preferred stimuli and the gains observed in Fig. 3.5 and Fig. 3.4c. Based on these differences in predictions, we conclude that the efficient coding hypothesis provides more parsimonious optimality principle for sensory and perceptual coding than the optimal discriminability hypothesis.

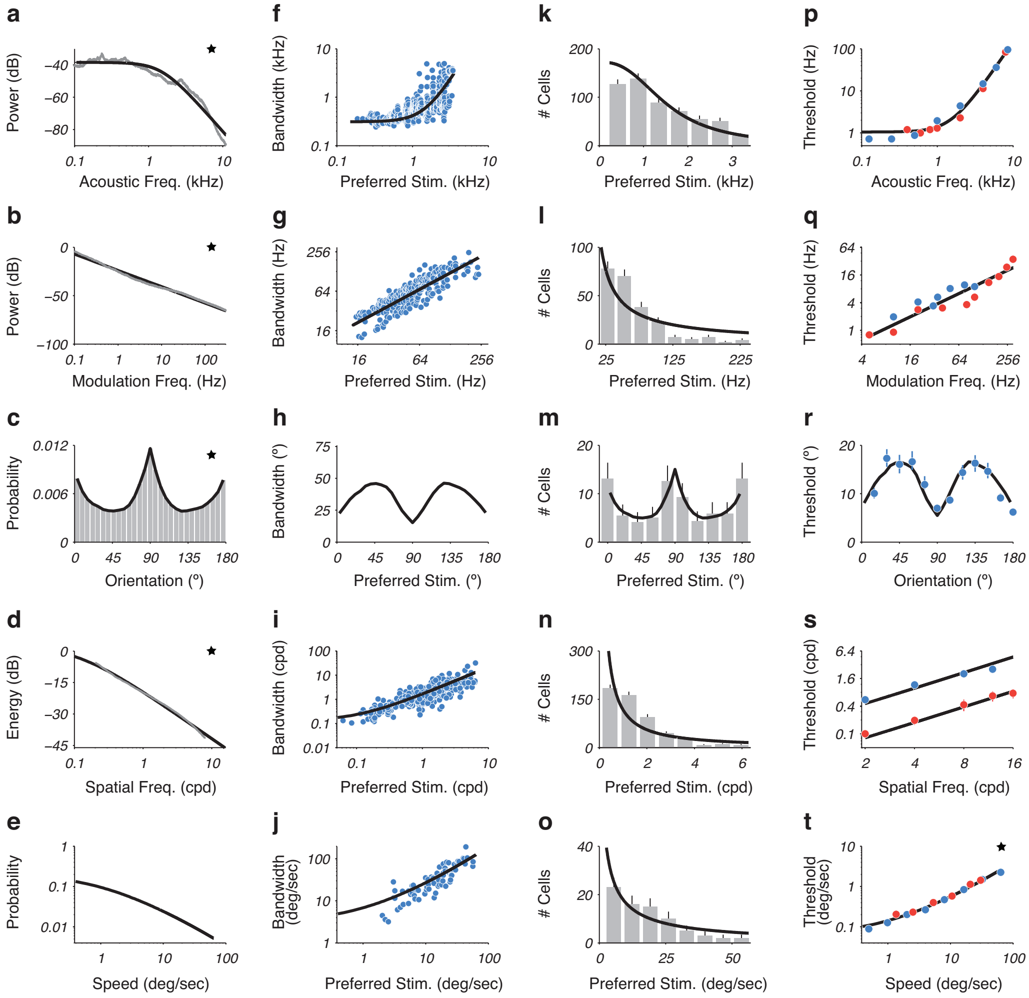

Figure 3.1: Testing the predicted relationship between sensory priors, neural population properties (tuning width and cell density) and psychophysical discrimination thresholds. Each row corresponds to a particular sensory attribute: acoustic frequency, modulation frequency, local orientation, spatial frequency, and speed. For each attribute, the data in the starred panel was fit with a parametric form or histogram density estimate (thick black lines, fitting details in methods). These curves were used (after transformation according to Eqs. (3.1) or (3.2)) to generate predictions for all other panels in the same row. Since the predictions include an unknown scale factor (that depends on resources N and R), each curve is rescaled to minimize the squared error to the associated data. **(a-e)** Estimated environmental distributions. Panels **a** and **b**, the distribution of acoustic and modulation frequencies were computed from commercially available compilations of animal vocalizations and background sounds [100–102]. Panel **c**, a histogram of

the distribution of local orientations in a large database of images containing both natural and man-made scenes [15, 36, 103, 104]. Panel **d**, distribution of spatial frequencies computed from the same image database. Panel **e**, theoretical prediction of the distribution of speeds encountered in the natural environment, based on the perceptual data in panel **t**, that is also consistent with the physiological data in panels **j** and **o**. See methods for details about how the distributions were estimated. **(f-j)** Physiologically measured tuning widths (as a function of preferred stimulus) of neural populations known to be tuned for each attribute. Panel **f**, data were obtained from 553 auditory nerve fibers measured in cats [105], adapted from [28]. Panel **g**, data were obtained from 262 neurons in the central nucleus of the inferior colliculous measured in cats [106]. Panel **h**, theoretical prediction of orientation tuning widths for a population of orientation-tuned V1 simple cells that is also consistent with the measured environmental distribution, cell density, and psycophysical thresholds from panels **c**, **m**, and **r**. Panel **i**, data were obtained from 538 cells in Macaque primary visual cortex (V1) [107]. Panel **j**, data were obtained from 76 speed-tuned cells in Macaque middle temporal cortex (MT) [108]. **(k-o)** Histograms of the number of cells tuned to each stimulus value. Panels **k**,**l**,**n**,&**o**, histograms were computed from the data in panels **f**,**g**,**i**,&**j**. Panel **m**, cell density reported for a population of 79 orientation-tuned V1 simple cells recorded foveally [109]. **(p-t)** Discrimination thresholds averaged across multiple human subjects for each sensory attribute. Panel **p**, acoustic frequency discrimination thresholds from two different studies shown in red [110] and blue [111]. Panel **q**, modulation frequency discrimination thresholds from two different studies shown in red [112] and blue [113]. Panel **r**, orientation discrimination thresholds [36]. Panel **s**, spatial frequency discrimination thresholds measured with sinusoidal gratings at 10% contrast (red) [114], and 25% contrast (blue) [115]. Two predictions are shown corresponding to different scale factors, reflecting the different stimulus conditions. Panel **t**, speed discrimination thresholds from two different studies shown in red [116] and blue [117].

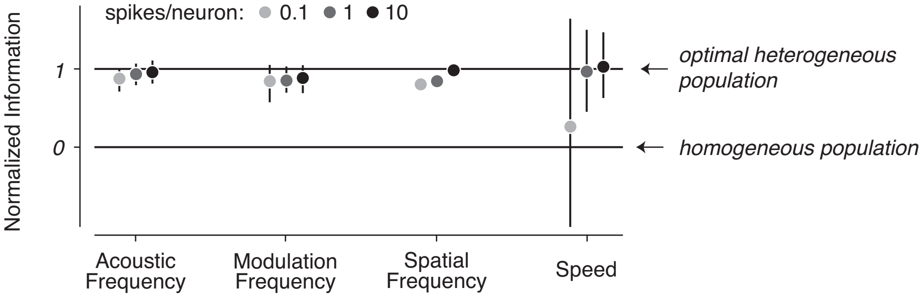

Figure 3.2: Normalized information transmitted by the observed neural populations for three different resource constraints. Error bars denote the 5th and 95th percentiles from 1000 bootstrap estimates. Values equal to 1 indicate information transmission of the optimal population parameterized by Eq. (3.1). Values equal to 0 indicate information transmission of the optimal homogeneous population.

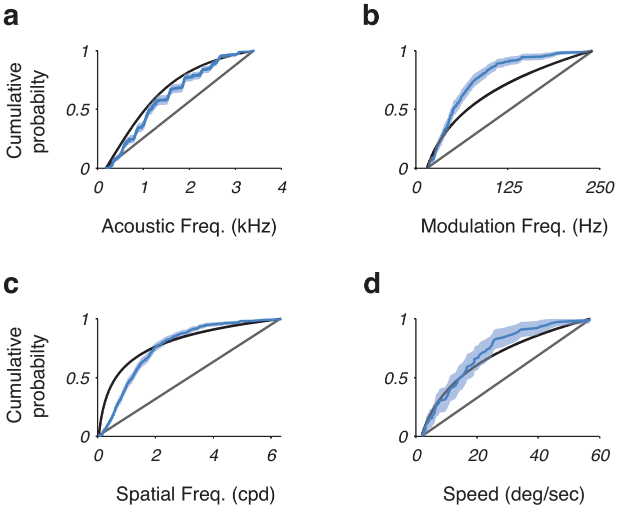

Figure 3.3: A comparison of the cumulative distribution of the estimated sensory prior (black line), and the empirical cumulative distribution of the cell density (blue line), for each attribute. Blue shaded regions denote 95% confidence intervals of the empirical cumulative distributions obtained from $1,000$ bootstrap estimates. Panel **a**, acoustic frequency. Panel **b**, modulation frequency. Panel **c**, spatial frequency. Panel **d**, speed.

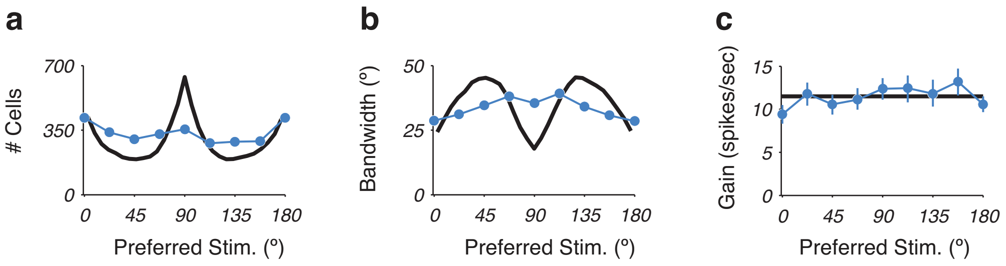

Figure 3.4: Orientation tuning preferences of $2,598$ simple cells in cat area 17 as reported in [132]. Predictions of the data based on the environmental distribution (black curves) overestimate the tuning anisotropies. Panel **a**, number of cells tuned to each orientation. The distribution shows clear peaks at horizontal and vertical orientations that are statistically significant ($p < 0.05, \chi^2$ test). Panel **b**, orientation tuning width as a function of preferred orientation. The cells show a significant narrowing of orientation tuning at horizontal orientations, but not to vertical orientations. Panel **c**, mean response amplitude (gain) as a function of preferred orientation. There is no systematic variation of gain as a function of stimulus value, in accordance with the predictions of our framework.

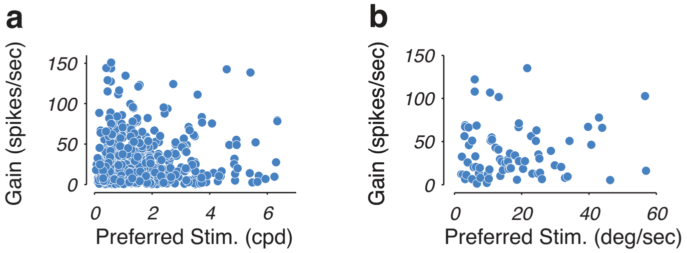

Figure 3.5: There is no systematic variation of cell gain (measured as the maximum average firing rate of each cell) with preferred stimuli. Panel **a**, data were obtained from 538 spatial frequency tuned cells in Macaque V1 (same data as in Fig. 3.1 **i,n**). Panel **b**, data were obtained from 76 speed tuned cells in MT (same data as in Fig. 3.1 **j,o**).

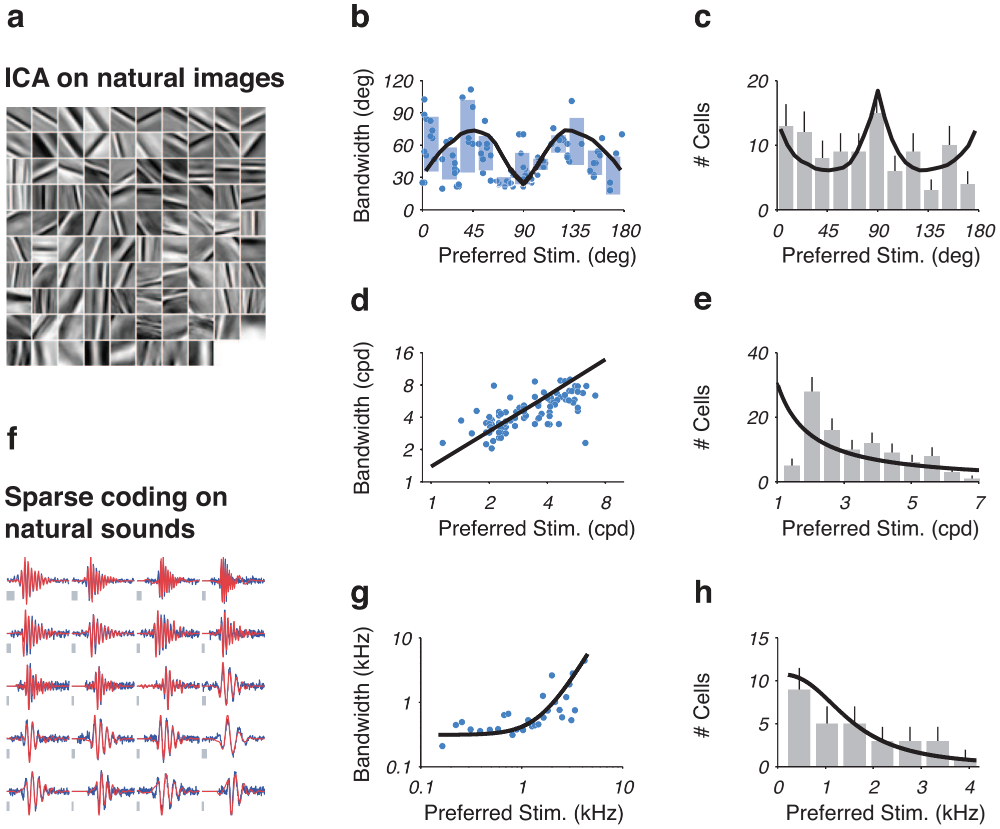

Figure 3.6: Panel **a**, linear receptive fields, derived from natural images using independent components analysis, resemble the receptive fields of V1 simple cells. (**b - e**) tuning characteristics of these receptive fields match the predictions of our efficient coding framework (thick black lines). Panel **b**, orientation tuning widths computed from the receptive fields in panel a. Blue bars show the mean and standard deviation of the bandwidths in 16.37 degree bins. Panel **c**, histogram density estimate of the preferred orientations of these receptive fields. Panel **d**, spatial frequency tuning widths computed from the receptive fields in panel **a**. Panel **e**, histogram density estimate of the preferred spatial frequencies of these receptive fields. Low spatial frequencies are underestimated due to the small patch sizes. Panel **f**, linear receptive fields, derived from natural sounds (red) using sparse coding, resemble the receptive fields of auditory nerve fibers (blue). Figure adapted from [28]. Panels **g** & **h**, tuning characteristics of these receptive fields match the predictions of our efficient coding framework (thick black lines). Panel **g**, acoustic frequency tuning widths adapted from [28]. Panel **h**, histogram density estimate of the preferred acoustic frequencies of the receptive fields from panel f.

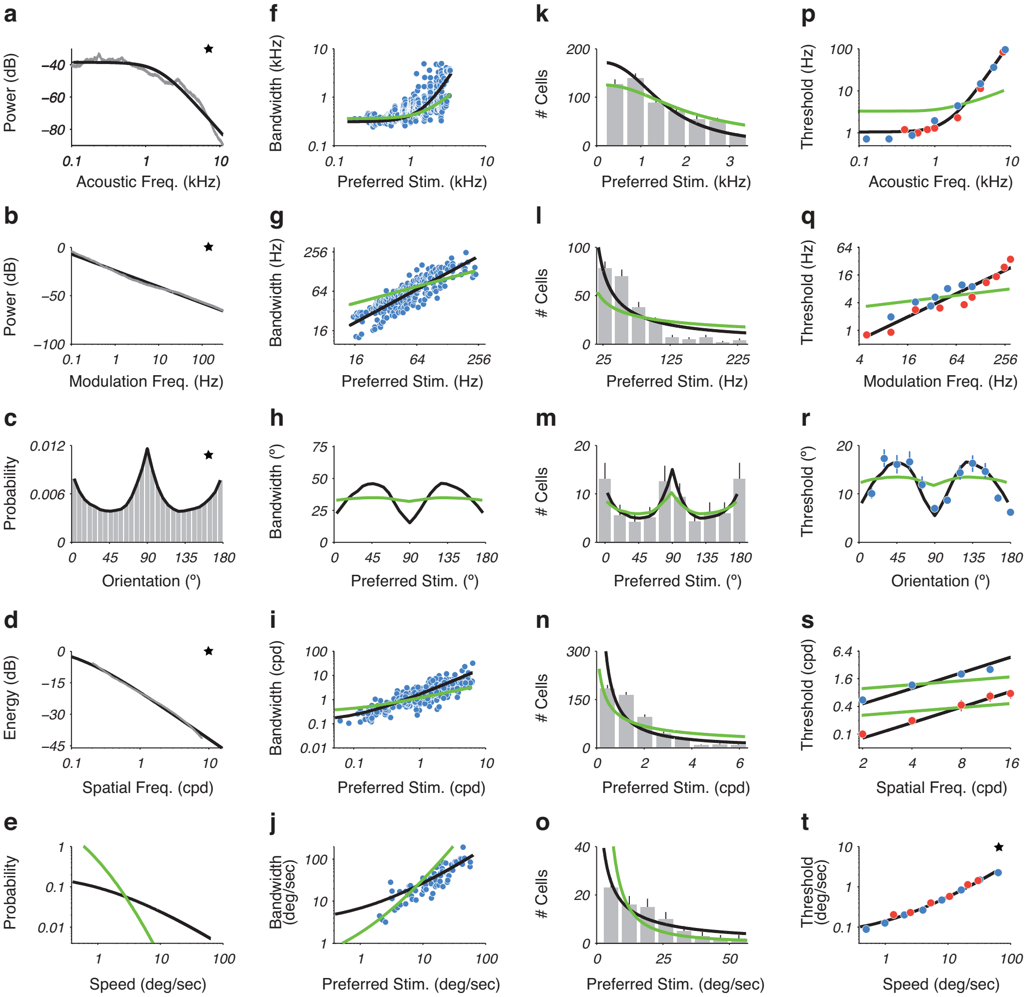

Figure 3.7: Comparing predicted relationships between sensory priors, neural population properties (tuning width and cell density) and psychophysical discrimination thresholds for the efficient coding hypothesis (thick black lines) and the optimal discrimination hypothesis (thick green lines). Each panel is the same as in Fig. 3.1, except here the curves fitted to the data in the starred panels were used (after transformation according to Eqs. (3.9) or (3.10)) to generate predictions generated from the optimal discriminability hypothesis Each curve is rescaled to minimize the squared error to the associated data.

Chapter 4

Neural Implementation of Bayesian Estimation using Efficient Population Codes

4.1 Introduction

Perception has been described as a process of inference, in which human observers combine their noisy sensory measurements and prior knowledge of the environment to construct an estimate of the physical world [4]. Bayesian statistics provides a powerful framework for understanding this inferential process with respect to three fundamental components [32]: (1) a likelihood function, which represents an observers degree of belief about an environmental signal based on sensory evidence; (2) a prior probability distribution, which characterizes an observers sense of the plausibility of different signal values in the environment; and (3) a loss function, which specifies the cost of making estimation errors. The central postulate in Bayesian statistics is that the prior and likelihood can be combined to form a posterior distribution, which reflects the plausibility of signal values arising from the environment based on the current sensory evidence and prior knowledge. As the uncertainty about the sensory evidence increases, the posterior rationally relies more heavily on the prior. An optimal point estimate of a signal value can be obtained by choosing the signal value that minimizes the expectation, over the posterior distribution, of the loss function. Experimental evidence suggests that human performance in perceptual [32, 34–36], sensorimotor, [37, 38], and cognitive [39] tasks is consistent with such optimal Bayesian estimation. Given the behavioral data in support of the Bayesian framework, a fundamental problem in neuroscience is to understand how the probability distributions required by the Bayesian machinery may be represented, learned, and computed with in the brain.

A substantial literature considers how a likelihood function can be represented by population of neurons, in which each neuron's mean response to a scalar variable is characterized by a tuning curve [e.g., 54–62]. In these models, sensory evidence corresponds to the noisy neural responses to the actual sensory input. When neural responses are Poisson distributed and statistically independent, the log-likelihood function can be explicitly computed as a sum (over neurons) of each neuron's log tuning curve, weighted by its response [59–61]. The uncertainty about sensory evidence (as characterized by the width of the likelihood function) increases with a decrease in the number of neurons, the total average firing rate, the inverse tuning widths, or some combination of these [78–81]. When the neuronal variability arises from an exponential family with linear sufficient statistics (which includes the independent Poisson model) recent work shows that adding the firing rates of two neural populations, each corresponding to a different sensory cue about a stimulus, implicitly multiplies the likelihoods they represent. This provides a plausible neural calculus for behavioral data that suggests human observers combine likelihood information in cue combination tasks [42].

The means by which prior information is represented and computed with in the brain is less well explored. In many cases, the prior is assumed to be uniform, and therefore the posterior distribution is simply proportional to the likelihood. However, a number of psychophysical studies suggest that human observers rely on non-uniform priors to produce (biased) estimates when sensory evidence is weak [e.g.,34–36]. One proposal for embedding prior information in neural responses, is to encode it in the same way as a likelihood [62]. In this case the log prior can be explicitly computed as firing rate weighted sum of the log tuning curves of a separate neural population. A desirable aspect of this approach is that a log posterior distribution can be implicitly computed by simply adding the firing rates of neurons representing the likelihood to the firing rates of neurons representing the prior, analogous to the case of combining likelihoods. However, the representation of the prior will be noisy due to neuronal variability, which may lead to suboptimal estimates of the true posterior distribution. An alternative approach interprets spontaneous neural activity as samples from a prior distribution defined over the latent variables of a generative model of the sensory input [63, 64]. Stimulus evoked activity is interpreted as samples from the posterior distribution over the latent variables. Although certain predictions of this approach are consistent with physiological data [64], it remains unclear how a stimulus estimate may be obtained from these responses, or how these models of neural processing are related to more conventional tuning curve based models of population coding.

A substantial body of work explores the neural implementation of estimators, or decoders, that can compute stimulus estimates based on noisy population responses [54–61, 136, 137]. These results all rely on choosing the stimulus value that maximizes the likelihood (ML estimate) or the posterior distribution (MAP estimate) of the stimulus values that generated the observed population response. Consider a downstream population implementing a MAP estimate from the neural responses of an encoding population, which is characterized by tuning curves and independent Poisson noise. The decoder must first explicitly compute a likelihood by computing a sum of tuning curves, each weighted by the firing rates of the encoding population, then add the result to a log prior distribution to obtain a log posterior distribution. The MAP estimate would then be the value associated with the decoder neuron with the maximum response, corresponding to a winner take all readout mechanism. Aspects of this implementation are undesirable [46]. First, the decoder requires a separate population of cells that has precise knowledge of the tuning curves in the encoding population. It is not clear how these tuning curves can be learned from the neural responses of the input population. Second, the decoder must have explicit knowledge of the prior distribution. If the prior is encoded in a separate population [62], then the decoder requires an additional population of cells that has full knowledge of the tuning curves representing the prior. Finally, the winner take all rule is highly sensitive to noise.

A more plausible decoder − the population vector (or center of mass decoder) − produces a stimulus estimate by computing a firing rate weighted average of the preferred stimulus values of each neuron. The population vector has a rich history in experimental neuroscience, where it has been used to accurately predict the direction of arm movements from neural responses measured in a variety of motor and premotor areas [69–72], the orientation of visual stimuli from responses in primary visual cortex [73], the direction of saccades from responses in the superior colliculus [74], and the position of a rat from the responses of place cells in the hippocampus [75]. Recent theoretical work has shown that the population vector can compute the mean of a posterior distribution [76, 77], which is optimal for minimizing a square loss function, and is known as a Bayes least squares estimator. These results rely on strong assumptions about the encoding population, which are asserted for the explicit purpose equating a Bayes least squares estimator to a population vector.

Here, we derive and examine precise conditions under which a Bayes least squares estimator can be implemented with a population vector like decoder. We assume sensory variables are encoded by a heterogeneous neural population with tuning curves that are optimized for the transmission of sensory information, subject to limitations on the number of neurons and the total average spike rate. This encoder implicitly represents the sensory prior in the distribution of preferred stimuli, is consistent with physiological data, and can predict perceptual discrimination thresholds of an optimal observer [99]. Given this encoder, we derive a novel decoder that can approximate the mean and variance of a posterior distribution. The decoder is based on a sampling approximation of the Bayes least squares estimator, where the preferred stimuli are construed as a set of samples from the prior. Similar to the population vector, it computes weighted averages of the preferred stimuli. However, the firing rates are not used directly as weights, but are first convolved with a linear filter then exponentiated. The decoder is neurally plausible, and requires knowledge only of the preferred stimuli and a fixed filter, and not the prior or tuning curves. Simulations demonstrate that it outperforms the standard population vector, and converges to the true Bayes least squares es-

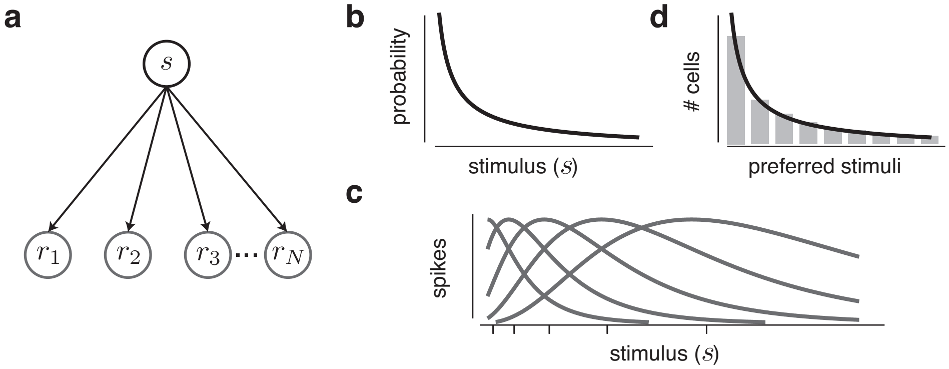

Figure 4.1: Generative model of the population response, and its neural representation with an efficient neural code. Panel **a**, a stimulus drawn from $p(s)$ generates a stochastic population response. The firing rate of the n^{th} neuron, r_n, is a Poisson random variable drawn from the probability distribution $p(r_n|s)$. The responses are independent conditioned on the stimulus value such that probability of a population response is $p(\vec{r}|s) = \prod_{n=1}^{N} p(r_n|s)$. Panel **b**, an example prior distribution from which an efficient population code can be constructed. Panel **c**, tuning curves denote the expected firing rate of each neuron in the efficient population, $h_n(s) = E_{r_n|s}(r_n)$, which correspond to the rate parameters of the Poisson distributions. The tuning curves are designed to maximize mutual information between the stimuli drawn from $p(s)$, and the responses. Tick marks denote the preferred stimulus value of each neuron. Panel **d**, the preferred stimuli represent samples from the prior distribution. A histogram of the preferred stimuli approximates the prior (thick black curve).

timator with as few as 10 neurons (or samples from the prior). We discuss how to test for signatures of this decoders use with psychophysical measurements of estimation bias.

4.2 Efficient encoding model

We assume a conventional model for a population of N neurons responding to a single scalar variable, s [54–62,136,137]. The number of spikes emitted (per unit time) by the nth neuron is a sample from an independent Poisson process, with mean rate determined by its tuning function, $h_n(s)$. The probability density of the population response can be written as

$$p(\vec{r}|s) = \prod_{n=1}^{N} \frac{h_n(s)^{r_n}\ e^{-h_n(s)}}{r_n!}. \tag{4.1}$$

We assume that stimulus values are drawn from a prior distribution, $p(s)$, which can have an arbitrary shape. A generative model of the population coding model is shown in Fig. 4.1a.

Most studies on population coding assume a convolutional population of neurons, where each neuron's tuning curve is a shifted copy of a prototype tuning curve, $h_n(s) = h(s - n)$ [e.g.,54–61, 78–81, 136, 137]. The tuning curve shape, $h(\cdot)$, is often assumed to be a Gaussian, raised cosine, or a Von Mises function. Here, we consider tuning curves that maximize the information transmitted about stimulus values drawn from $p(s)$, subject to constraints on the number of neurons, N, and the total number of spikes R. The optimal tuning curves may be obtained by warping the tuning curves in the standard convolutional code by the cumulative distribution of the prior ($P(s) = \int_{-\infty}^{s} p(t)\,\mathrm{d}t$) and scaling the entire population to fire a maximum of R spikes [99]. As a result, each optimized tuning curve can be expressed as, $h_n(s) = Rh(s - NP(s))$. An example of an efficient population code for an exponential prior distribution is shown in Fig. 4.1b-c.

The efficient population code implicitly embeds the prior distribution in the preferred stimuli and widths of the tuning curves. The preferred stimulus value of each neuron is $s_n = NP^{-1}(n)$, where n is an integer ranging from 1 to N. These values represent samples from the prior distribution (Fig. 4.1d). [1] The tuning width of each neuron in the population is inversely proportional to the prior. This can be seen by taking a first-order Taylor expansion of $P(s)$ around s_n to obtain $h_n(s) \approx Rh(Np(s_n)(s - s_n))$.

4.3 Bayesian estimation

We seek a decoder that can appropriately leverage the prior information embedded in the tuning curves of the efficient encoding population to produce an optimal Bayesian perceptual estimate. An estimator, $\hat{s}(\vec{r}(s))$, is a deterministic function that takes as input a noisy population response to a sensory parameter, $\vec{r}(s)$, and outputs an estimate of that parameter $\hat{s}$. Due to the response variability, the estimated values can differ for repeated presentations of the same sensory stimulus (Fig. 4.2). For notational convenience, we drop the explicit dependence of the firing rates on the unobserved stimulus such that $\vec{r}(s) = \vec{r}$.

The first step in a Bayesian estimation strategy is to construct a posterior probability distribution of the possible stimuli that gave rise to the observed population response,

$$p(s|\vec{r}) = \frac{p(\vec{r}|s)p(s)}{p(\vec{r})}.$$

Here, the likelihood, $p(\vec{r}|s)$, is a function of s evaluated for a single observation of $\vec{r}$. The denominator, $p(\vec{r}) = \int p(\vec{r}|s)p(s)\,\mathrm{d}s$, is a normalizing constant that ensures

[1] The standard (stochastic) way to obtain N samples from $p(s)$ is to generate a N random draws from a uniform distribution between 1 and N, and then pass these through the inverse cumulative distribution [129]. Repeating this procedure will generate a different set of N samples each time the procedure is repeated. The efficient coding formulation offers a *deterministic* set of samples that does not change unless the prior changes.

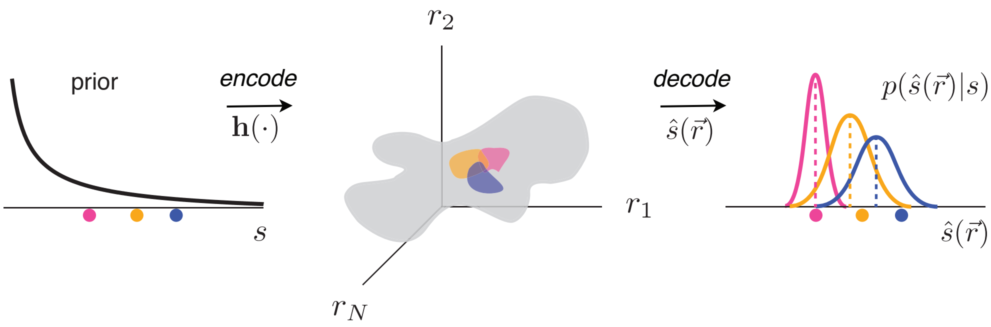

Figure 4.2: Illustration of the encoding decoding process. Stimuli drawn from the prior distribution (colored dots) are encoded by a population of neurons, characterized by tuning curves $\mathbf{h}(\cdot)$. The encoding process is noisy. As a result, repeated presentations of the same stimulus value, yield a distribution of N dimensional neural responses (clouds colored corresponding to the stimuli that caused them). A deterministic estimator decodes a single pattern of activity into an estimate $\hat{s}(\vec{r})$. Due to the response noise, the estimated values for particular stimulus condition are also random variables that can be modeled with a probability distribution $p(\hat{s}(\vec{r})|s)$ (colored curves). The difference between the true stimulus and the average estimates (dashed lines) represents the estimation bias. The width of these distributions corresponds to the estimation variance.

the posterior is a proper probability distribution.

An optimal Bayesian estimator computes an estimate from the posterior distribution that minimizes a loss function, $L(s, \hat{s}\,(\vec{r}))$, which characterizes the average cost of a mismatch between estimated stimulus and the true value:

$$\hat{s}_{\text{OPT}}\,(\vec{r}) = \underset{\hat{s}(\vec{r})}{\arg\min} \int p(s|\vec{r}) L(s, \hat{s}\,(\vec{r}))\,\mathrm{d}s. \tag{4.2}$$

A common loss function is the squared error,

$$L(s, \hat{s}\,(\vec{r})) = (s - \hat{s}\,(\vec{r}))^2, \tag{4.3}$$

which can be expressed in terms of two quantities: the bias, $b(s)$, and the variance $\sigma^2(s)$ (Fig. 4.2). The bias is the difference between the average of $\hat{s}(\vec{r})$ across trials that use the stimulus s and the true value of s:

$$b(s) = E_{R|S}\left[\hat{s}(\vec{r})\right] - s \tag{4.4}$$

An estimator is termed unbiased if $b(s) = 0$ for all stimulus values. Estimation biases can arise from the influence of the prior (when the sensory evidence is weak), the loss function, or even the likelihood. The variance of the estimator quantifies how much the estimate varies about its mean value:

$$\sigma^2(s) = E_{R|S}\left[\left(\hat{s}(\vec{r}) - E_{R|S}\left[\hat{s}(\vec{r})\right]\right)^2\right]. \tag{4.5}$$

The bias and variance can be used to compute the trial averaged conditional mean squared estimation (MSE) error:

$$\mathrm{MSE}(s) = E_{R|S}\left[\left(\hat{s}(\vec{r}) - s\right)^2\right] = b^2(s) + \sigma^2(s). \tag{4.6}$$

The total mean squared estimation error is simply the expectation of the conditional error under the prior distribution: $\mathrm{MSE} = \int \mathrm{MSE}(s)p(s)\,\mathrm{d}s$.

The Bayes least squares (BLS) estimator can be derived from Equations. 4.2 & 4.3 as computing the mean of the posterior distribution:

$$\hat{s}_{\mathrm{BLS}}(\vec{r}) = E_{S|R}[s] = \int sp(s|\vec{r})\,\mathrm{d}s. \tag{4.7}$$

An analytical expression for this estimator does not exist for our encoding model.

4.4 An importance sampling approximation to Bayesian estimation

A common technique to evaluate expectations, like the one necessary for computing the BLS estimator (Eq. 4.7), is to approximate them with a sum such that,

$$\int sp(s|\vec{r})\,\mathrm{d}s \approx \frac{1}{L}\sum_{l=1}^{L} s_l p(s_l|\vec{r}), \tag{4.8}$$

where $s_l \sim p(s|\vec{r})$ represents one of L samples from the posterior probability distribution. The approximation converges to the true expectation as the number of samples tends to infinity. For the brain to use this trick, it would need access to samples from the posterior distribution, but the efficient code only produces samples from the *prior* distribution in the form of preferred stimuli, $s_n \sim p(s)$. A well known sampling method, called importance sampling, can correctly use this embedded prior information to approximate the Bayes least squares estimator.

In general form, an importance sampler approximates the expected value of any function, $f(s)$, with samples from a proposal distribution, $q(s)$, that is easy to obtain samples from,

$$E_{S|R}\left[f(s)\right] = \int f(s)\frac{p(s|\vec{r})}{q(s)}q(s)\,\mathrm{d}s \approx \frac{1}{L}\sum_{l=1}^{L} f(s_l)\frac{p(s_l|\vec{r})}{q(s_l)}. \tag{4.9}$$

Here, s_l represents one of L samples from the proposal distribution ($s_l \sim q(s)$). If the proposal distribution is chosen to be the prior distribution, and the preferred stimuli are used as samples from this prior, then the importance sampler (Eq. 4.9) can be expressed (using Bayes rule) as,

$$E_{S|R}\left[f(s)\right] \approx \frac{1}{N}\sum_{n=1}^{N} f(s_n)\frac{p(s_n|\vec{r})}{p(s_n)} = \frac{\frac{1}{N}\sum_{n=1}^{N} f(s_n)p(\vec{r}|s_n)}{p(\vec{r})}. \tag{4.10}$$

To compute the marginal probability of the observed population response, $p(\vec{r})$, we again approximate the necessary integral over the joint distribution, using the preferred stimuli as samples from the prior: $p(\vec{r}) = \int p(\vec{r}|s)p(s)\,\mathrm{d}s \approx \frac{1}{N}\sum_{n=1}^{N} p(\vec{r}|s_n)$. Substituting this expression into Eq. 4.10, and choosing $f(s) = s$, allows us to express the Bayes least squares estimator as,

$$\hat{s}_{\mathrm{BLS}}\left(\vec{r}\right) \approx \frac{\sum_{n=1}^{N} s_n p(\vec{r}|s_n)}{\sum_{n=1}^{N} p(\vec{r}|s_n)}. \tag{4.11}$$

4.5 The Bayesian population vector

The importance sampling approximation to the Bayes least squares estimator exhibits a striking similarity to a standard population vector decoder. The population vector (PV) computes a firing rate weighted average of the preferred stimuli of the cells, and can be expressed as,

$$\hat{s}_{\mathrm{PV}}\left(\vec{r}\right) = \frac{\sum_{n=1}^{N} s_n r_n}{\sum_{n=1}^{N} r_n}. \tag{4.12}$$

By inspection, if one assumes $r_n \propto p(\vec{r}|s_n)$, then $\hat{s}_{\mathrm{PV}}\left(\vec{r}\right) = \hat{s}_{\mathrm{BLS}}\left(\vec{r}\right)$ [76, 77]. However, this assumption is both artificial and clearly violated by our response model.

To derive a version of the importance sampling approximation of the BLS estimator that does not rely on this assumption, we expand out the likelihood weights, $p(\vec{r}|s_n)$ according to Eq. 4.1, and substitute them into Eq. 4.11 to obtain,

$$\begin{aligned}\hat{s}_{\text{BLS}}\left(\vec{r}\right) &\approx \frac{\sum_{n=1}^{N} s_n \exp\left(\sum_{m=1}^{N} r_m \log h_m(s_n) - \sum_{m=1}^{N} h_m(s_n) - \sum_{m=1}^{N} \log(r_m!)\right)}{\sum_{n=1}^{N} \exp\left(\sum_{m=1}^{N} r_m \log h_m(s_n) - \sum_{m=1}^{N} h_m(s_n) - \sum_{m=1}^{N} \log(r_m!)\right)} \\ &= \frac{\sum_{n=1}^{N} s_n \exp\left(\sum_{m=1}^{N} r_m \log h_m(s_n)\right)}{\sum_{n=1}^{N} \exp\left(\sum_{m=1}^{N} r_m \log h_m(s_n)\right)} \qquad (4.13)\end{aligned}$$

In the second step, we use the fact that for an efficient population code, $\sum_{m=1}^{N} h_m(s_n) = R$, and can thus be pulled out of the sum over n in both numerator and denominator to cancel. The fact that the tuning curves sum to a constant has previously been assumed to be true [61], but here we have derived it from the fundamental principle of efficient coding [99]. The term $\sum_{m=1}^{N} \log(r_m!)$ does not depend on n, and therefore also cancels out in the numerator and denominator.

The term $h_m(s_n)$ represents the average response of the m^{th} neuron to the stimulus preferences of all the other neurons in the population. In a convolutional population code, this set of weights is the same for all m neurons, since each tuning curve overlaps the same amount with its neighbors. This is also true for the tuning curves in the efficient population code, as they are obtained from warping a convolutional code by the cumulative of the prior distribution, ensuring that the original overlap between tuning curves is preserved. As a result, the term $\sum_{m=1}^{N} r_m \log h_m(s_n)$ can be expressed as a convolution of the neural responses with a fixed linear filter, $\vec{w}$, which depends on the (log of the) amount of overlap between tuning curves. Incorporating this into Eq. 4.13, we obtain an expression for the importance sampling approximation to the Bayes least squares estimator, which we term the Bayesian population vector (BPV):

$$\hat{s}_{\text{BPV}}\left(\vec{r}\right) = \frac{\sum_{n=1}^{N} s_n \exp\left(\sum_{m=1}^{N} r_m \vec{w}_{m-n}\right)}{\sum_{n=1}^{N} \exp\left(\sum_{m=1}^{N} r_m \vec{w}_{m-n}\right)} \tag{4.14}$$

The BPV can be directly mapped onto a compact, biologically plausible neural circuit (Fig. 4.3). To produce a Bayes least squares estimate, a downstream neural population receives inputs (in the form of spike counts) from the responses of the efficient population. Each downstream neuron linearly pools together the responses of neurons in the input population that have similar stimulus preferences. The result is then passed through a static exponential non-linearity. This simple linear-nonlinear (LN) cascade corresponds to approximating the likelihood, or sensory evidence. The output of this cascade is then sent to a standard population vector decoder, which makes use of the embedded prior information to compute the posterior mean.

4.6 Simulations

The Bayesian population vector provides a remarkably accurate approximation to the omniscient Bayes least squares estimator (which has explicit access to the prior and likelihood), in terms of mean squared error, across a wide range of N and R (Fig. 4.4a). We computed the mean squared errors of these estimators, based on the responses of a population code optimized for transmitting information about samples from an exponential prior distribution, as shown in Fig. 4.1. To compute the mean squared error, we first simulated neural population responses to $1,000$ repeats of each stimulus value. The response of each neuron to a single stimulus value corresponds to a sample from a Poisson distribution, with the rate parameter determined by the neuron's tuning curve, evaluated at the stimulus value. From these neural responses, we computed stimulus estimates, using the omniscient Bayes least squares estimator (Eq. 4.7) and the Bayesian population vector (Eq. 4.14). The sample mean and variance of these estimates, conditioned on the stimulus condition, produces estimates of the estimation bias (Eq. 4.4) and variance (Eq. 4.5) respectively. To compute the mean squared error, we first approximate conditional mean squared error from the bias and variance terms (Eq. 4.6), then compute its expectation under the sensory prior by numerical quadrature.

The mean squared error of the Bayesian population vector converges to that of the Bayes least squares estimator as the number of neurons (or samples from the prior) increases, independent of the total average firing rates. In a low firing rate regime (0.1 maximum average spikes/neuron) the approximation is to within 1% of the true MSE with as few as 10 neurons. However, when the same 10 neurons are firing a maximum of 100 spikes each, the mean squared error of the BPV is 25% larger than that of the BLS estimator. In this regime, the likelihood is very narrow (due to the abundance of spikes) relative to the spacing of the preferred stimuli (which is large due to the small number of neurons). As a result modified weights, given by the continuous likelihood function evaluated at the preferred stimuli, $p(\vec{r}|s_n)$, may reduce to a delta function centered on the single preferred stimulus value that lies in the non-zero support of the likelihood. When the weights correspond to a delta function, the BPV reduces to a winner take all readout mechanism which is suboptimal relative to the omniscient BLS estimator operating in the same metabolic regime.

The standard population vector provides an accurate approximation to the

Bayes least square estimator, in terms of mean squared error, in a low firing rate regime (0.1 maximum average spikes/neuron), and a very poor approximation when the neurons fire 1 or 10 spikes per neuron (Fig. 4.4b). In these higher firing rate regimes, the population vector becomes increasingly suboptimal (by orders of magnitude) as the number of neurons increases. To further understand the differences in mean squared error between these three estimators we examine their bias and variance properties (Fig. 4.5). We find that the Bayesian population vector has similar bias and variance properties to the Bayes least squares estimator. Both of these estimators become unbiased as the uncertainty about the sensory evidence decreases (either with an increase in N, R, or both). However, the population vector does not take likelihood information into account, and is therefore biased by the prior even when sensory evidence is strong. This suggests that the standard population vector is insufficient for explaining psychophysical evidence that suggests human observers properly take sensory uncertainty into account when making perceptual estimates.

4.7 Experimental predictions

The simplest physiological prediction of the BPV decoder is that the preferred stimuli of the tuning curves in a brain region thought to encode a particular sensory attribute, should represent samples from the prior probability of occurrence of that attribute. Testing other aspects of the BPV with physiological measurements is difficult, as it requires directly determining neurons that are responsible for producing estimates, characterizing their response properties (as a function of the spike rates of the encoding population), and comparing them to the response properties of the hypothetical decoder neurons.

Perceptually, the BPV makes a testable prediction about estimation bias. As the simulations show, the decoder bias depends on the observer's sensory prior, and the number of neurons and total average firing rates used in a task (Fig. 4.5b,e). These biases can be measured perceptually for a human observer and compared to those of the BPV with the parameters N and R fit to the data. The simplest way to measure estimation bias is to give a subject a tool to indicate the perceived value of a stimulus parameter. For example, a probe stimulus shown at low signal to noise ratio (SNR), s_L, could be briefly presented, followed by a high SNR test stimulus, s_H, that the subject could adjust until they perceived the two stimuli to be equal. The bias in estimating the probe would then be, $b(s_L) = E[s_H] - s_L$, where the expectation is taken across trials that used the same probe stimulus. As the SNR of s_L decreases, we assume that R decreases, thus broadening the observers likelihood function such that their estimates will be more readily biased by the sensory prior. We also assume that the SNR of s_H is high such that the internal representation of test stimulus is unbiased. This method of estimating perceptual bias may be corrupted by additional motor biases involved in manipulating the tool to adjust the test stimulus.

Alternatively, perceptual bias can be measured in a two alternative forced choice (2AFC) paradigm where, on a given trial, subjects are presented with both high and low SNR stimuli, and asked to report a binary decision about whether $s_L > s_H$. The stimulus conditions under which the subject makes the correct choice 50% of the time corresponds to the point of subjective equality, in which the internal representations of the two stimulus values, s_L and s_H, are the same. If the subject is using the BPV decoder, this implies

$$E_{R|S}\left[\hat{s}_{\mathrm{BPV}}(\vec{r}(s_L))\right] = E_{R|S}\left[\hat{s}_{\mathrm{BPV}}(\vec{r}(s_H))\right]$$
$$b(s_L) + s_L = b(s_H) + s_H.$$

If the subjects estimates of the high SNR stimulus are unbiased, then perceptual bias for the low SNR stimulus conditions is $b(s_L) = s_H - s_L$. Although this experimental paradigm eliminates potential motor biases, it requires more trials than the estimation task to estimate perceptual biases. Nevertheless, in both cases, evidence in support of the use of the BPV in a perceptual task may be obtained if experimental measurements of perceptual biases match those of the biases exhibited by the BPV for stimuli drawn from a known sensory prior.

4.8 Discussion

There is considerable evidence that human judgements of many perceptual attributes can be described by Bayesian observer models, in which noisy sensory measurements are combined with long term prior knowledge to obtain stimulus estimates. We have elucidated a physiological substrate for this probabilistic computation. To do so, we model the noisy measurement process by a population of neurons, each tuned to particular values of the sensory attribute, such that the tuning curves are optimized for transmitting information about stimulus values drawn from the prior distribution. Such an efficient code implicitly embeds the prior distribution in the density of tuning curves, and allows for the expression of a population vector like decoder that closely approximates an explicit Bayesian estimator that is optimal for minimizing a square loss function.

Our basic formalism of characterizing a sensory measurement with the responses of an encoding population, followed by an estimate produced by a decoding population, is highly oversimplified. Sensory computations occur in cascades, with each layer adding additional noise. Designating a single neural population as an encoder or decoder is thus somewhat artificial. Furthermore, many estimation problems require maintaining and propagating an estimate of variable that evolves over time. We have effectively ignored the time course of neuronal responses, which can be used to produce and propagate stimulus estimates and their corresponding uncertainties [138,139]. We have also assumed that neuronal responses are independent, conditioned on the stimulus value. Neural populations do exhibit correlated variability [86,87], and possibly higher order dependencies. A downstream decoder must have full knowledge of these dependencies in order to compute an optimal estimate; however, increased attention in a task can decrease correlated variability [140], which may help justify the use of a decoder that assumes independent responses. Despite these caveats, our results present an important step forward in the population coding literature to understand how the brain how the brain might represent and compute with the probabilities required by the Bayesian machinery.

In our framework, we assume the encoding of long term prior information in the structure of the tuning curves. As a result, the prior may induce biases in the *likelihood* function. To gain an intuition for how this may occur, we express an analytical solution for a maximum likelihood (ML) estimator by assuming that the optimally efficient tuning curves can be approximated by Gaussian functions with mean parameters s_n and standard deviation parameters σ_n. The resulting

ML estimator [56, 141, 142],

$$\hat{s}_{\mathrm{ML}}\left(\vec{r}\right) \approx \frac{\sum_{n=1}^{N} s_n \frac{r_n}{\sigma_n^2}}{\sum_{n=1}^{N} \frac{r_n}{\sigma_n^2}},$$

is also similar to a population vector. In addition to the preferred stimuli corresponding to samples from the prior, the tuning widths are inversely proportional to the prior, $\frac{1}{\sigma_n} \propto p(s_n)$. As a result, spikes produced from neurons with narrower tuning curves will bias the peak of the likelihood towards their preferred stimuli stronger than spikes from neurons with wider tuning curves. As a result, it may be possible that perceptual biases can be explained by a maximum likelihood estimation strategy operating on the responses of an efficient population code.

Although we have described a decoder than can explicitly produce stimulus estimates from a posterior distribution, in some cases it is desirable for the brain to leave the representation probabilistic in order to facilitate further computations [46, 62]. Our framework can be extended to do so, by implicitly encoding a Gaussian approximation to the posterior in terms of its mean and variance parameters. The mean of the Gaussian approximation to the posterior, μ, is approximated by the output of the Bayesian population vector (Eq. 4.14). The variance parameter approximation, σ^2, can be approximated from the same importance sampling approximation (Eq. 4.9, with $f(s) = s^2$) as

$$\begin{aligned}\sigma^2 &= E_{S|R}\left[s^2\right] - E_{S|R}^2\left[s\right] \\ &\approx \frac{\sum_{n=1}^{N} s_n^2 \exp\left(\sum_{m=1}^{N} r_m \vec{w}_{m-n}\right)}{\sum_{n=1}^{N} \exp\left(\sum_{m=1}^{N} r_m \vec{w}_{m-n}\right)} - \left(\frac{\sum_{n=1}^{N} s_n \exp\left(\sum_{m=1}^{N} r_m \vec{w}_{m-n}\right)}{\sum_{n=1}^{N} \exp\left(\sum_{m=1}^{N} r_m \vec{w}_{m-n}\right)}\right)^2\end{aligned}$$

Computing the variance of the posterior distribution requires the same circuit as computing the mean, except with additional quadratic non-linearities. Such Gaussian approximations to the posterior are common in variational approaches to solving inference problems in machine learning [129], and in some cases it has been shown that human observers typically only represent the mean and variance of probability distributions [143]. We are currently developing novel perceptual tasks to determine what types of distributions the brain chooses to encode, and what tradeoffs are involved in learning complex distributions [134].

The Bayesian population vector consists of a cascade of three canonical neural computations, linear filtering, exponentiation, and divisive normalization. Each of these have considerable physiological support across a wide range of sensory systems [144, 145]. Linear non-linear (LN) computations provide an accurate description of the functional response properties of neurons in early sensory systems [146, 147]. In these models, the linear filter is usually applied directly to the stimulus variable, but here we apply it to spiking responses of an input population. The output of the LN cascade undergoes divisive normalization, which has been implicated in numerous theoretical goals of neural coding [144]. A small subset of these goals include producing invariance with respect to particular stimulus dimensions [148], aiding the discrimination of stimuli by a downstream population [149], reducing the statistical independence between neural responses (a noiseless formulation of the efficient coding hypothesis) [150], and marginalizing probability distributions [151]. Furthermore, cascades of LN processing, followed by divisive normalization, have been proposed in the computer vision community as optimal artificial neural network architectures for object recognition [152]. We add optimal Bayesian estimation to the long list of rationales for these canonical computations. Our results support the idea that the brain relies on a small set computational modules, that are cascaded across brain regions and sensory modalities, to solve a diverse array of computational problems.

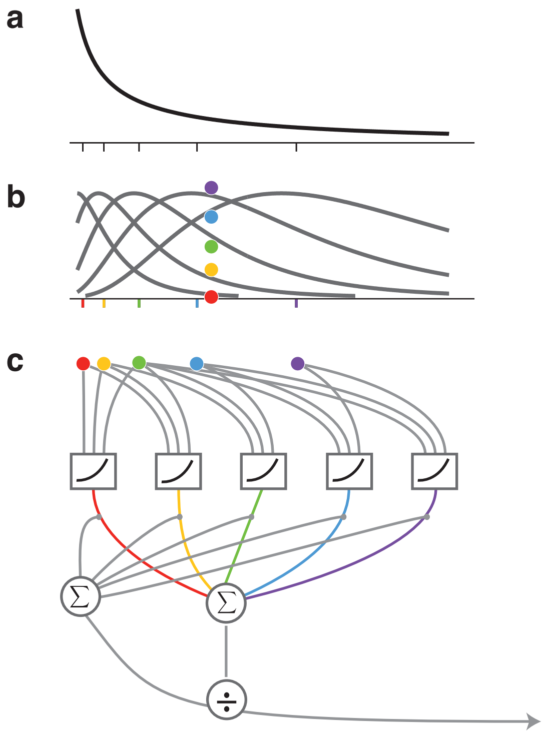

Figure 4.3: Neural implementation of the Bayesian population vector. Panel **a**, example of a prior distribution over the stimulus variable. Panel **b**, population code designed to maximize information transmission. Tick marks denote the preferred stimuli, s_n, of each neuron which also represent samples from the prior: $s_n \sim p(s)$. These are also shown as tick marks on the stimulus axis of the prior. The firing rate of each neuron in response to a particular stimulus value is represented by a dot color coded by the preferred stimulus of the corresponding tuning curve. Panel **c**, to produce a Bayesian perceptual estimate of the stimulus value that generated the observed response, the firing rate of each cell is first convolved with a linear filter. The filter pools together the firing rates of neurons with similar stimulus preferences (triplets of thin gray lines). The output of this convolution is then exponentiated (boxes with non-linearities) and passed to a standard population vector decoder. The population vector weights the convolved and exponentiated firing rates by the preferred stimulus of each cell (thick lines color coded by the stimulus preferences of the encoder neurons). The result is summed and divisively normalized by the unweighted sum of the modified firing rates. The result is the Bayes least square estimate of the stimulus that produced the observed population response.

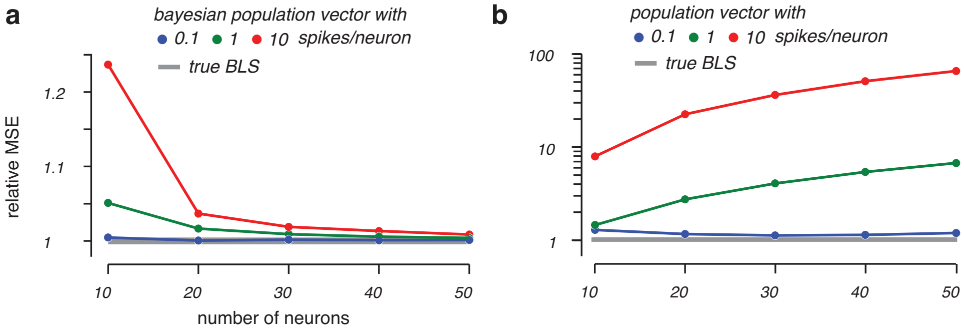

Figure 4.4: Mean squared estimation error of the Bayesian population vector decoder and the standard population vector decoder, relative to the omniscient Bayes least squares decoder for a variety of resource constraints. A relative MSE value of 1 indicates a perfect approximation to the BLS. Values greater than 1 indicate a poor approximation. Panel **a**, the Bayesian population vector accurately approximates the omniscient Bayes least squares estimator (in terms of mean squared error) over a wide range of resource constraints. Panel **b**, the standard population vector is suboptimal in most cases.

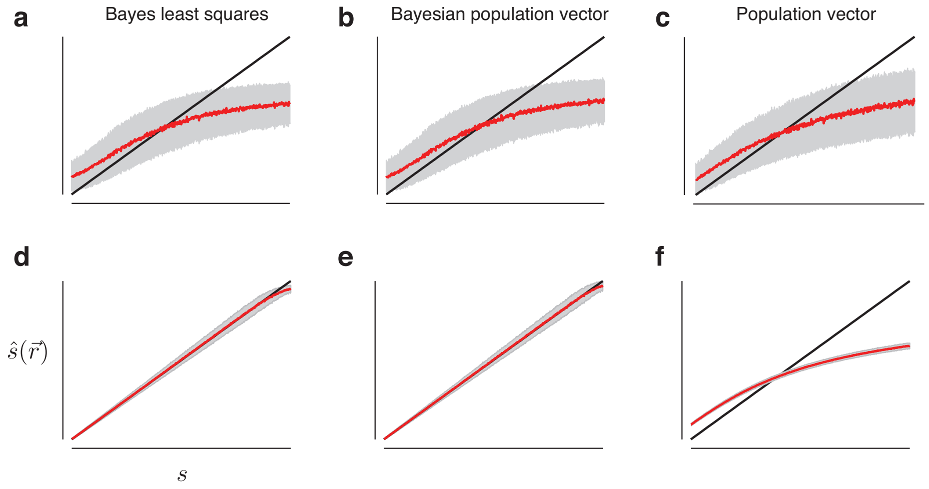

Figure 4.5: Bias and variance properties of the Bayes least squares estimator, Bayesian population vector, and population vector. Each column corresponds to a different estimator. Each row corresponds to different resource constraints on the encoding population. The expected value of the estimates is shown in red. The estimation bias is the deviation of this expected value from the actual stimulus (black). Estimation variance is shown as a grey shaded region about the mean estimates. Panels **a-c**, estimates obtained from the responses of an encoding population with $N = 20$ neurons with a maximum average firing rate of 1 spike for each neuron. In this regime, all estimators have similar bias and variance properties. Panel **d-f**, estimates obtained from the responses of an encoding population with $N = 20$ neurons with a maximum average firing rate of 100 spikes for each neuron. In this regime, the BLS and BPV estimators become unbiased, but the PV remains biased.

Chapter 5

General Discussion

A fundamental goal in sensory neuroscience is to understand the transformations of natural signals into neural representations that enable perception. We have explored this problem in the context of two well known theories – the efficient coding hypothesis [1, 2], and perception as optimal Bayesian estimation [32]. In particular, we developed a variant of the efficient coding hypothesis which led to the prediction that heterogeneities in neural tuning and perceptual discriminability arise because of heterogeneities in the prior probability of encountering different sensory inputs in the natural environment. We supported this prediction with environmental, physiological, and perceptual data obtained for two auditory and three visual attributes. We then modeled perceptual estimates as arising from an efficient encoder - Bayesian decoder cascade. We developed a neurally plausible Bayesian decoder that extracts the embedded prior information in the efficient encoder, and combines it with likelihood information to produce a perceptual estimate. The decoder can be implemented in a LN cascade with divisive normalization, and closely approximates an ideal Bayesian observer model. Our results establish a tight link between efficient coding and Bayesian estimation, and suggest how to relate both ideas directly to data from physiological and perceptual experiments. There are several intriguing opportunities for future research.

Our framework examines the neural and perceptual coding of sensory inputs whose probability of occurrence in the environment is stable over long time scales. However, both perception, and the response properties of neurons are known to adapt to changes in sensory inputs that occur over the timescale of milliseconds to minutes (see [153, 154] for review). For example, prolonged exposure to an oriented pattern causes the perception of a subsequent test pattern to be biased away from the adapted pattern (for small differences between the test and adaptor) [155]. Similar repulsive biases have been reported after adaptation to the direction of motion stimuli [156]. For both types of stimuli, discrimination thresholds measured for test stimuli near the adaptor either decrease modestly [157–159], or remain unchanged [160, 161]. Investigations into the neural correlates of these perceptual effects have focused on examining adaptation induced changes in the tuning properties of V1 neurons responding to oriented stimuli, and MT neurons responding to motion stimuli. In V1, adaptation causes a decrease in the gain of neurons tuned to the adaptor, and a repulsive shift in the stimulus preferences of neurons whose tuning curves flank the adaptor [162]. In MT, adaptation also causes a decrease in the gain of cells tuned to the adaptor [163], but causes an *attractive* shift in the stimulus preferences of tuning curves that flank the adaptor [164]. Furthermore, the flanking tuning curves exhibit narrower tuning widths after adaptation. Taken together, these results suggests that cortical areas may differ in their adaptation properties.

Aspects of the physiological and perceptual consequences of adaptation are inconsistent with efficient coding and Bayesian estimation. If we interpret adaptation as a means by which a neural population updates its sensory prior (by maximizing information transmission) then the population should allocate more neurons with narrower tuning widths to the adapted stimulus condition, and maintain a constant gain across all neurons. This solution is inconsistent with the physiological data showing a decrease in the gains of cells tuned to the adaptor, and in some cases a decrease in the number of cells tuned to the adaptor. Furthermore, updating the prior to reflect the boost in occurrence of the adapted stimulus would lead an optimal Bayesian estimator to exhibit *attractive* estimation biases, which is at odds with the perceptually documented repulsive biases [165]. To examine whether a temporarily suboptimal (with respect to information transmission) coding scheme could explain adaptation, we considered what would happen if we re-optimized the gain for the updated sensory prior while holding the cell density fixed. In this case, the optimal gain is again constant, which implies that the neurons tuned to the adaptor will fire (on average) *more* spikes, simply because the adaptor occurs more frequently and the tuning widths no longer equalize the responses. The solution also implies that discriminability will not be affected by adaptation. Neither of these predictions are consistent with the data. A possible explanation for the repulsive perceptual biases seen during adaptation, is that the decoder is “unaware” of the adaptation induced changes in the encoder [96]. Such suboptimal decoding can qualitatively account for repulsive biases, and in our framework, corresponds to the Bayesian decoder not updating its pooling weights or preferred stimuli post adaptation. It seems as though adaptation cannot be described by the kinds of optimality principles explored here. Developing a complete theoretical explanation of adaptation provides an important opportunity for future work.

Given the physiological changes observed during adaptation, it may be possible to develop an unsupervised learning rule (one that does not have direct access to the sensory prior or samples from the sensory prior) that arrives at our efficient coding solution without the explicit computation and maximization of mutual information. There are three main signals the brain could monitor and adjust over different time scales to achieve this. First, the solution requires that the mean response of each neuron (averaged across all possible stimuli) should be the same. Simple forms of adaptation achieve this over relatively fast time scales, which suggests that the brain has the capacity to regulate average firing rates dynamically with changes in sensory inputs, perhaps for maintaining homeostasis. Second, the solution requires that the tuning widths should be inversely proportional to the sensory prior. Such changes in tuning widths could arise from sparsity constraints, which put pressure on a neural system to minimize the number of active neurons [27, 28]. For example, if a neuron is too active (its preferred stimulus occurs frequently), then it can narrow its tuning curve to be less selective, which would effectively decrease the number of times it responds to stimuli. Finally, the solution requires that the neuronal tuning curves evenly tile the stimulus space. In order to achieve this, neurons need to coordinate with their nearest neighbors, perhaps through local recurrent connectivity, such that their responses are not overly redundant. Understanding the mechanisms by which these three signals, mean firing rate, sparsity, and redundancy, could be monitored and adjusted to arrive at an efficient coding solution offers an important opportunity for future research.

Once the efficient code is learned, it may also be possible to learn the properties of the down stream neurons responsible for approximating Bayes least squares estimation. Assuming the exponential non-linearity is either known a priori or fundamentally constrained by the neurobiology, the remaining parameters of the Bayesian population vector are the linear pooling weights, and the preferred stimuli of the input population. The pooling weights represent the amount of overlap between neighboring tuning curves. This knowledge could be embedded in the same local recurrent connectivity that monitors redundancy to enforce tiling in the efficient input layer. Once the pooling weights are learned, the preferred stimuli could be learned from labeled data (pairs of stimuli and their associated neural responses after pooling, exponentiation, and normalization) through linear regression. In an unsupervised setting, recent work proposes an online learning rule for a Bayes least squares estimator [166]. The learning rule requires knowledge of the probability distribution of measurements (in our case this is the marginal distribution of responses, $p(\vec{r})$) and the details of the likelihood function. Given that the brain has unlimited access to spikes, it seems reasonable to assume that it can appropriately estimate the marginal distribution over spike counts. Whether the brain can combine this information with knowledge of its noise processes to learn the parameters of the Bayesian population vector provides an intriguing opportunity for future research.

In this thesis, we have provided a theoretical framework for understanding how sensory signals get transformed into neural representations that lead to perception, and suggested how to test the framework with environmental, physiological, and perceptual data. The theories presented here offer only a small glimpse into the potentially vast set of computational principles underlying the neural code. Given recent advances in neural recording technologies and statistical methodologies, the time seems ripe for developing and testing a new generation of theories that can further unlock the mysteries of the brain.

Bibliography

[1] F Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.

[2] H Barlow. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234, 1961.

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

[4] H Helmholtz. *Treatise on physiological optics*. Thoemmes Press, Bristol, UK, 2000.

[5] EP Simoncelli and BA Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.

[6] SB Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift Für Naturforschung.*, 36(9-10):910–912, 1981.

[7] T von der Twer and DI MacLeod. Optimal nonlinear codes for the perception of natural colours. *Network*, 12(3):395–407, August 2001.

[8] MD McDonnell and NG Stocks. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Physical Review Letters*, 101(5):58103, 2008.

[9] N Brenner, W Bialek, and R de Ruyter van Steveninck. Adaptive rescaling maximizes information transmission. *Neuron*, 26(3):695–702, June 2000.

[10] AL Fairhall, GD Lewen, W Bialek, and RR de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

[11] JP Nadal and N Parga. Non linear neurons in the low noise limit: A factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 1994.

[12] AJ Bell and TJ Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[13] H Barlow. Redundancy reduction revisited. *Network*, 12(3):241–253, August 2001.

[14] AJ Bell and TJ Sejnowski. The ”independent components” of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, December 1997.

[15] JH van Hateren and A van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 265(1394):359–366, March 1998.

[16] AA Faisal, LPJ Selen, and DM Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, April 2008.

[17] JJ Atick and AN Redlich. Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320, 1990.

[18] JJ Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 22(1-4):4–44, 1992.

[19] MV Srinivasan, SB Laughlin, and A Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, 216(1205):427–459, November 1982.

[20] MJ Wainwright. Visual adaptation as optimal information transmission. *Vision research*, 39(23):3960–3974, November 1999.

[21] Y Karklin and EP Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In *Adv. Neural Information Processing Systems (NIPS*11)*, volume 24, Cambridge, MA, May 2012. MIT Press. Presented at Neural Information Processing Systems 24, Granada, Spain, Dec 13-15 2011.

[22] SB Laughlin, R de Ruyter van Steveninck, and John C. Anderson. The metabolic cost of neural information. *Nature Neuroscience*, 1(1):36–41, May 1998.

[23] D Attwell and SB Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of cerebral blood flow and metabolism: official*

journal of the International Society of Cerebral Blood Flow and Metabolism, 21(10):1133–1145, October 2001.

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

[25] WB Levy and RA Baxter. Energy efficient neural codes. *Neural Computation*, 8(3):531–543, 1996.

[26] V Balasubramanian and MJ Berry. A test of metabolically efficient coding in the retina. *Network (Bristol, England)*, 13(4):531–552, November 2002.

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

[28] EC Smith and MS Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, February 2006.

[29] DL Ringach. Spatial structure and symmetry of Simple-Cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88(1):455–463, July 2002.

[30] WS Geisler, J Najemnik, and AD Ing. Optimal stimulus encoders for natural tasks. *Journal of Vision*, 9(13):17.1–16, 2009.

[31] ET Jaynes. How does the brain do plausible reasoning? technical report 421. In GJ Erickson and CR Smitt, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering.*, volume 1, pages 1–23. 1988, Kluwer Academic Publishers, 1957.

[32] DC Knill and W Richards. *Perception as Bayesian inference.* Cambridge University Press, Cambridge, UK, 1996.

[33] K Doya, S Ishii, RPN Rao, and A Pouget, editors. *Bayesian Brain: Probabilistic Approaches to Neural Coding.* MIT Press, January 2007.

[34] Y Weiss, EP Simoncelli, and EH Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604, June 2002.

[35] AA Stocker and EP Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585, April 2006.

[36] AR Girshick, MS Landy, and EP Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932, June 2011.

[37] KP Körding and DM Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, January 2004.

[38] H Tassinari, TE Hudson, and MS Landy. Combining priors and noisy visual cues in a rapid pointing task. *The Journal of Neuroscience*, 26(40):10154–10163, October 2006.

[39] TL Griffiths, C Kemp, and JB Tenenbaum. *Bayesian models of cognition.* Cambridge Handbook of Computational Psychology. Cambridge, UK, 2008.

[40] LY Maloney. Statistical decision theory and biological vision. In Dieter Heyer Ph D. essor of Psychology and Rainerusfeldessor of Cognitive Science,

editors, *Perception and the Physical World*, page 145–189. John Wiley & Sons, Ltd, 2002.

[41] P Mamassian, MS Landy, LT Maloney, R Rao, B Olshausen, and M Lewicki. Bayesian modelling of visual perception. In *Probabilistic Models of the Brain: Perception and Neural Function*, pages 13–36. MIT Press, 2002.

[42] MO Ernst and HH Bülthoff. Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–169, April 2004. PMID: 15050512.

[43] D Kersten, P Mamassian, and A Yuille. Object perception as bayesian inference. *Annual review of psychology*, 55:271–304, 2004. PMID: 14744217.

[44] K Körding. Decision theory: What "Should" the nervous system do? *Science*, 318(5850):606–610, October 2007.

[45] L Paninski and EP Simoncelli. Nonparametric inference of prior probabilities from bayes-optimal behavior. In Y Weiss, B Scholkopf, and J Platt, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 18, pages 1067–1074, Salt Lake City, Utah, 2006.

[46] EP Simoncelli. Optimal estimation in sensory systems. In M Gazzaniga, editor, *The Cognitive Neurosciences, IV*, pages 525–535. MIT Press, October 2009.

[47] DG Peli and B Farell. Why use noise? *16*, pages 647–653, 1999.

[48] MS Landy, R Goutcher, J Trommershäuser, and P Mamassian. Visual estimation under risk. *Journal of Vision*, 7(6), April 2007.

[49] J Trommershäuser, LT Maloney, and MS Landy. Decision making, movement planning, and statistical decision theory. *Trends in cognitive sciences*, 12(8):291–297, August 2008.

[50] G E Hinton. How neural networks learn from experience. *Scientific American*, 267(3):144–151, September 1992. PMID: 1502516.

[51] EP Simoncelli. *Distributed Analysis and Representation of Visual Motion*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, January 1993. Also available as MIT Media Laboratory Vision and Modeling Technical Report #209.

[52] CH Anderson and DC Van Essen. Neurobiological computational systems. New York, 1994. IEEE Press.

[53] MJ Barber, JW Clark, and CH Anderson. Neural representation of probabilistic information. *Neural Comput.*, 15(8):1843–1864, August 2003.

[54] HS Seung and H Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90(22):10749–10753, November 1993.

[55] E Salinas and LF Abbott. Vector reconstruction from firing rates. *Journal of Computational Neuroscience*, 1(1-2):89–107, June 1994. PMID: 8792227.

[56] HP Snippe. Parameter extraction from population codes: A critical assessment. *Neural Computation*, 8(3):511–529, 1996.

[57] TD Sanger. Probability density estimation for the interpretation of neural population codes. *J Neurophysiol*, 76(4):2790–2793, October 1996.

[58] K Zhang, I Ginzburg, BL McNaughton, and TJ Sejnowski. Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, 79(2):1017–1044, February 1998.

[59] RS Zemel, P Dayan, and A Pouget. Probabilistic interpretation of population codes. *Neural Computation*, 10(2):403–430, February 1998.

[60] A Pouget, P Dayan, and RS Zemel. Inference and computation with population codes. *Annu Rev Neurosci*, 26:381–410, 2003.

[61] M Jazayeri and JA Movshon. Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5):690–696, 2006.

[62] WJ Ma, JM Beck, PE Latham, and A Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, October 2006.

[63] Patrik O. Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. In *Advances in Neural Information Processing Systems*, page 2002. MIT Press, 2002.

[64] P Berkes, G Orbán, M Lengyel, and J Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, January 2011.

[65] SJ Gershman, E Vul, and JB Tenenbaum. Multistability and perceptual inference. *Neural computation*, 24(1):1–24, January 2012.

[66] Paul W. Glimcher and Aldo Rustichini. Neuroeconomics: The consilience of brain and decision. *Science*, 306(5695):447–452, October 2004.

[67] W Schultz. Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current opinion in neurobiology*, 14(2):139–147, April 2004.

[68] LP Sugrue, GS Corrado, and WT Newsome. Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, 6(5):363–375, May 2005.

[69] AP Georgopoulos, AB Schwartz, and RE Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, September 1986.

[70] R Caminiti, PB Johnson, C Galli, S Ferraina, and Y Burnod. Making arm movements within different parts of space: the premotor and motor cortical representation of a coordinate system for reaching to visual targets. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 11(5):1182–1197, May 1991.

[71] JF Kalaska, R Caminiti, and AP Georgopoulos. Cortical mechanisms related to the direction of two-dimensional arm movements: relations in parietal area 5 and comparison with motor cortex. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 51(2):247–260, 1983.

[72] PA Fortier, JF Kalaska, and AM Smith. Cerebellar neuronal activity related to whole-arm reaching movements in the monkey. *Journal of neurophysiology*, 62(1):198–211, July 1989. PMID: 2754472.

[73] CD Gilbert and TN Wiesel. The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat. *Vision Research*, 30(11):1689–1701, 1990.

[74] C Lee, WH Rohrer, and DL Sparks. Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, 332(6162):357–360, March 1988. PMID: 3352733.

[75] MA Wilson and BL McNaughton. Dynamics of the hippocampal ensemble code for space. *Science*, 261(5124):1055–1058, August 1993.

[76] L Shi and TL Griffiths. Neural implementation of hierarchical bayesian inference by importance sampling. In *NIPS*, pages 1669–1677, 2009.

[77] BJ Fischer and JL Peña. Owl's behavior and neural representation predicted by bayesian inference. *Nature Neuroscience*, 14(8):1061–1066, August 2011.

[78] K Zhang and TJ Sejnowski. Neuronal tuning: To sharpen or broaden? *Neural Computation*, 11(1):75–84, January 1999.

[79] A Pouget, S Deneve, JC Ducom, and PE Latham. Narrow versus wide tuning curves: What's best for a population code? *Neural Computation*, 11(1):85–90, January 1999.

[80] WM Brown and A Bäcker. Optimal neuronal tuning for finite stimulus spaces. *Neural Computation*, 18(7):1511–1526, July 2006.

[81] MA Montemurro and S Panzeri. Optimal tuning widths in population coding of periodic variables. *Neural Computation*, 18(7):1555–1576, July 2006.

[82] A Gersho and RM Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, 1991.

[83] N Brunel and JP Nadal. Mutual information, fisher information, and population coding. *Neural Computation*, 10(7):1731–1757, September 1998.

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

[85] KH Britten, MN Shadlen, WT Newsome, and JA Movshon. Responses of neurons in macaque MT to stochastic motion signals. *Visual neuroscience*, 10(6):1157–1169, December 1993.

[86] E Zohary, MN Shadlen, and WT Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, July 1994. PMID: 8022482.

[87] A Kohn and MA Smith. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *The Journal of Neuroscience*, 25(14):3661–3673, April 2005.

[88] JM Beck, WJ Ma, PE Latham, and A Pouget. Probabilistic population codes and the exponential family of distributions. *Prog Brain Res*, 165:509–519, 2007.

[89] D Cox and D Hinkley. *Theoretical statistics*. London: Chapman and Hall., 1974.

[90] M Shamir and H Sompolinsky. Implications of neuronal diversity on population coding. *Neural Computation*, 18(8):1951–1986, 2006.

[91] LF Abbott and P Dayan. The effect of correlated variability on the accuracy of a population code. *Neural Computation*, 11(1):91–101, January 1999.

[92] P Seriès, PE Latham, and A Pouget. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature neuroscience*, 7(10):1129–1135, October 2004.

[93] I Dean, NS Harper, and D McAlpine. Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, 8(12):1684–1689, 2005.

[94] S Durant, CWG Clifford, NA Crowder, NSC Price, and MR Ibbotson. Characterizing contrast adaptation in a population of cat primary visual cortical neurons using fisher information. *Journal of the Optical Society of America*, 24(6):1529–1537, June 2007.

[95] DA Gutnisky and V Dragoi. Adaptive coding of visual information in neural populations. *Nature*, 452(7184):220–224, March 2008.

[96] P Seriès, AA Stocker, and Simoncelli EP. Is the homunculus “aware” of sensory adaptation? *Neural Computation*, 21(12):3271–3304, September 2009.

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

[98] P Berens, S Gerwinn, A Ecker, and M Bethge. Neurometric function analysis of population codes. In *Advances in Neural Information Processing Systems 22*, page 90–98, 2009.

[99] D Ganguli and EP Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. In J Lafferty, CKI Williams, J Shawe-Taylor, RS Zemel, and A Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 23, pages 658–666, Salt Lake City, Utah, 2010.

[100] J Storm. Great smokey mountains national park: winter and spring. in: The macaulay library of natural sounds. ithaca, NY: cornell laboratory of ornithology., 1994.

[101] J Storm. Great smokey mountains national park: summer and fall. in: The macaulay library of natural sounds. ithaca, NY: cornell laboratory of ornithology., 1994.

[102] LH Emmons, BM Whitney, and DL Ross. Sounds of neotropical rainforrest mammals. in: The macaulay library of natural sounds. ithaca, NY: cornell laboratory of ornithology., 1997.

[103] E Doi, T Inui, TW Lee, T Wachtler, and TJ Sejnowski. Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Comput*, 15(2):397–417, February 2003.

[104] A Olmos and FAA Kingdom. *McGill Calibrated Image Database, http://tabby.vision.mcgill.ca*. 2004.

[105] LH Carney, MJ McDuffy, and I Shekhter. Frequency glides in the impulse responses of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 105(4):2384, 1999.

[106] FA Rodríguez, C Chen, HL Read, and Escabí MA. Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. *The Journal of Neuroscience*, 30(47):15969 –15980, November 2010.

[107] JR Cavanaugh, W Bair, and JA Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2530 –2546, November 2002.

[108] JA Movshon. Unpublished observations.

[109] RJW Mansfield. Neural basis of orientation perception in primate vision. *Science*, 186(4169):1133 –1135, December 1974.

[110] BCJ Moore. Frequency difference limens for short-duration tones. *The Journal of the Acoustical Society of America*, 54(3):610, 1973.

[111] CC Wier. Frequency discrimination as a function of frequency and sensation level. *The Journal of the Acoustical Society of America*, 61(1):178, 1977.

[112] C Formby. Differential sensitivity to tonal frequency and to the rate of amplitude modulation of broadband noise by normally hearing listeners. *The Journal of the Acoustical Society of America*, 78(1):70, 1985.

[113] J Lemanska, AP Sek, and Skrodzka EB. Discrimination of the amplitude modulation rate. *Archives of Acoustics*, 27(1):3–21, 2002.

[114] T Caeli, H Brettel, I Rentschler, and R Hilz. Discrimination thresholds in the two-dimensional spatial frequency domain. *Vision Research*, 23(2):129–133, 1983.

[115] D Regan, S Bartol, TI Beverly, and TJ Murray. Spatial frequency discrimination in normal vision and in patients with multiple sclerosis. *Brain*, 105(4):735 –754, December 1982.

[116] SP McKee and K Nakayama. The detection of motion in the peripheral visual field. *Vision Research*, 24(1):25–32, 1984.

[117] B De Bruyn and GA Orban. Human velocity and direction discrimination measured with random dot patterns. *Vision Research*, 28(12):1323–1335, 1988.

[118] P Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *Audio and Electroacoustics, IEEE Transactions on*, 15(2):70 – 73, June 1967.

[119] H Attias and CH Schreiner. Temporal Low-Order statistics of natural sounds. *NIPS*, 9:27—33, 1997.

[120] JH McDermott and EP Simoncelli. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5):926–940, September 2011.

[121] BR Glasberg and BC Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, August 1990.

[122] P Burt and E Adelson. The laplacian pyramid as a compact image code. *Communications, IEEE Transactions on*, 31(4):532 – 540, April 1983.

[123] H Farid and EP Simoncelli. Differentiation of discrete multidimensional signals. *Image Processing, IEEE Transactions on*, 13(4):496 –508, April 2004.

[124] GH Granlund and H Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, Norwell, Massachusetts, 1995.

[125] DJ Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, December 1987.

[126] DL Ruderman and W Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, August 1994.

[127] DW Dong and JJ Atick. Statistics of natural Time-Varying images. *Network: Computation in Neural Systems*, 6:345—358, 1995.

[128] S Roth and MJ Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1):33–50, January 2007.

[129] CM Bishop. *Pattern Recognition And Machine Learning*. Springer, August 2006.

[130] H Nover, CH Anderson, and GC DeAngelis. A logarithmic, Scale-Invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. *The Journal of Neuroscience*, 25(43):10049–10060, October 2005.

[131] S Appelle. Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, 78(4):266–278, 1972.

[132] L Baowang, MR Peterson, and RD Freeman. Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology*, 90(1):204 –217, July 2003.

[133] CA Rothkopf, TH Weisswange, and J Triesch. Learning independent causes in natural images explains the spacevariant oblique effect. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pages 1 –6, June 2009.

[134] J Freeman, GJ Brouwer, DJ Heeger, and EP Merriam. Orientation decoding depends on maps, not columns. *The Journal of Neuroscience*, 31(13):4792–4804, March 2011.

[135] A Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626 –634, May 1999.

[136] P Foldiak, FH Eeckman, and J Bower. The ideal homunculus: Statistical inference from neural population responses. In *Computation and Neural Systems*, pages 55–60. Kluwer, Norwell, Massachusetts, 1993.

[137] TD Sanger. Neural population codes. *Current Opinion in Neurobiology*, 13(2):238–249, April 2003.

[138] RPN Rao and DH Ballard. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9:721–763, 1997.

[139] S Deneve. Bayesian spiking neurons i: Inference. *Neural Computation*, 20(1):91–117, 2007.

[140] MR Cohen and JHR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–1600, December 2009.

[141] P Dayan and LF Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, December 2001. Published: Hardcover.

[142] WJ Ma. Signal detection theory, uncertainty, and poisson-like population codes. *Vision research*, 50(22):2308–2319, October 2010.

[143] AN Sanborn, TL Griffiths, and RM Shiffrin. Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, 60(2):63–106, March 2010.

[144] M Carandini and DJ Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, January 2012.

[145] M Kouh and T Poggio. A canonical neural circuit for cortical nonlinear operations. *Neural Computation*, 20(6):1427–1451, 2008.

[146] EJ Chichilnisky. A simple white noise analysis of neuronal light responses. *Network*, 12(2):199–213, May 2001.

[147] EP Simoncelli, JP Paninski, J Pillow, and O Schwartz. Characterization of neural responses with stochastic stimuli. *The Cognitive Neurosciences*, 2004.

[148] SR Olsen, V Bhandawat, and RI Wilson. Divisive normalization in olfactory population codes. *Neuron*, 66(2):287–299, April 2010.

[149] DL Ringach. Population coding under normalization. *Vision research*, 50(22):2223–2232, October 2010.

[150] O Schwartz and EP Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, August 2001.

[151] JM Beck, PE Latham, and A Pouget. Marginalization in neural circuits with divisive normalization. *The Journal of Neuroscience*, 31(43):15310–15319, October 2011.

[152] K Jarrett, K Kavukcuoglu, MA Ranzato, and Y LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146 –2153, October 2009.

[153] CWG Clifford. Perceptual adaptation: motion parallels orientation. *Trends in cognitive sciences*, 6(3):136–143, March 2002.

[154] Adam Kohn. Visual adaptation: Physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, 97(5):3155–3164, May 2007.

[155] JJ Gibson and M Radner. Adaptation, aftereffect, and contrast in the perception of tilted lines. i. quantitative studies. *Journal of Experimental Psychology*, 20:453–467, 1937.

[156] E Levinson and R Sekuler. Adaptation alters perceived direction of motion. *Vision research*, 16(7):779–781, 1976.

[157] D Regan and KI Beverley. Postadaptation orientation discrimination. *Journal of the Optical Society of America. A, Optics and image science*, 2(2):147–155, February 1985.

[158] CW Clifford, AM Wyatt, DH Arnold, ST Smith, and P Wenderoth. Orthogonal adaptation improves orientation discrimination. *Vision research*, 41(2):151–159, January 2001.

[159] RE Phinney, C Bowd, and R Patterson. Direction-selective coding of stereoscopic (cyclopean) motion. *Vision research*, 37(7):865–869, April 1997.

[160] HB Barlow, DIA Macleod, and van Meeteren. Adaptation to gratings: No compensatory advantages found. *Vision Research*, 16(10):1043–1045, 1976.

[161] K Hol and S Treue. Different populations of neurons contribute to the detection and discrimination of visual motion. *Vision research*, 41(6):685–689, March 2001. PMID: 11248258.

[162] SG Marlin, SJ Hasan, and MS Cynader. Direction-selective adaptation in simple and complex cells in cat striate cortex. *Journal of Neurophysiology*, 59(4):1314–1330, April 1988.

[163] A Kohn and JA Movshon. Neuronal adaptation to visual motion in area MT of the macaque. *Neuron*, 39(4):681–691, August 2003.

[164] A Kohn and JA Movshon. Adaptation changes the direction tuning of macaque MT neurons. *Nature neuroscience*, 7(7):764–772, July 2004. PMID: 15195097.

[165] AA Stocker and EP Simoncelli. Sensory adaptation within a bayesian framework for perception. In Y. Weiss, B Schölkopf, and B Platt, editors, *Adv. Neural Information Processing Systems (NIPS*05)*, volume 18, page 1291–1298, Cambridge, MA, May 2006. MIT Press.

[166] M. Raphan and E. P. Simoncelli. Least squares estimation without priors or supervision. *Neural Computation*, 23(2):374–420, February 2011. Published online, Nov 2010.