OPTIMAL NEURAL CODES FOR NATURAL STIMULI

Zhuo Wang

A DISSERTATION

in

Department of Mathematics

For the Graduate Group in Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

Daniel D. Lee, Professor, Electrical and Systems Engineering

Graduate Group Chairperson

Charles L. Epstein, Thomas A. Scott Professor, Mathematics

Dissertation Committee

Charles L. Epstein, Thomas A. Scott Professor, Department of Mathematics

Alan A. Stocker, Assistant Professor, Department of Psychology

Daniel D. Lee, Professor, Department of Electrical and System Engineering

OPTIMAL NEURAL CODES FOR NATURAL STIMULI

Dedicated to my wife and my parents.

ACKNOWLEDGEMENT

The PhD study is the most challenging and rewarding period of my life so far. Here I would like to thank my colleagues, family, friends and all those who, in one way or another, helped me to make it to the point of today. Without their help, this journey of PhD study would be a much more difficult one that I could not imagine.

First and foremost, I would like to thank my advisor Dan Lee, for his whole-hearted support, care, encouragement and inspiration throughout these years. There are lots of things that I will keep benefiting for my lifetime, both academically and life-wise – his vision on research directions, elegant execution when solving problems, and his resolute attitude towards his own life and career. I'm also very grateful to Alan Stocker, who always gives me inspiring comments on manuscripts that we have submitted together, who taught me the art of writing and communication. I also wish to acknowledge the help provided by Charles Epstein, whose dedication to our academic program is the key of every students' successful career. It is my greatest honor and pleasure to welcome Dan, Alan and Charles to be on my thesis committee.

I would like to mention my collaborators who shared the road with me to becoming a theoretical neuroscientist: Kevin Shi, Jerome Tubiana and Xue-xin Wei. My special thanks are extended to all current and former lab members from Dan's group and Alan's group. At the group meetings and journal clubs, there are always something to learn.

Last but not least, I want to thank my whole family, especially my wife Congyun, for all the love and support.

ABSTRACT

OPTIMAL NEURAL CODES FOR NATURAL STIMULI

Zhuo Wang

Daniel D. Lee

The efficient coding hypothesis assumes that biological sensory systems use neural codes that are optimized to best possibly represent the stimuli that occur in their environment. When formulating such optimization problem of neural codes, two key components must be considered. The first is what types of constraints the neural codes must satisfy? The second is the objective function itself – what is the goal of the neural codes? We seek to provide a systematic framework to address these types of problem.

Previous work often assume one specific set of constraint and analytically or numerically solve the optimization problem. Here we want to put everything in a unified framework and show that these results can be understood from a much more generalized perspective. In particular, we provide analytical solutions for a variety of neural noise models and two types of constraint: a range constraint which specifies the max/min neural activity and a metabolic constraint which upper bounds the mean neural activity.

In terms of objective functions, most common models rely on information theoretic measures, whereas alternative formulations propose incorporating downstream decoding performance. We systematically evaluate different optimality criteria based upon the L_p reconstruction error of the maximum likelihood decoder. This parametric family of optimal criteria includes special cases such as the information maximization criterion and the mean squared loss minimization of decoding error. We analytically derive the optimal tuning curve of a single neuron in terms of the reconstruction error norm p to encode natural stimuli with an arbitrary input distribution.

Under our framework, we can try to answer questions such as what is the objective function the neural code is actually using? Under what constraints can the predicted results provide a better fit for the actual data? Using different combination of objective function and constraints, we tested our analytical predictions against previously measured characteristics of some early visual systems found in biology. We find solutions under the metabolic constraint and low values of p provides a better fit for physiology data on early visual perception systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

Animals interact with their surrounding world on a daily basis and they perceive stimulus from the environment to ensure their survival. It is both appealing and crucial for the brains to find good representations of these stimulus inputs for advantages in surviving. To understand why the sensory information is encoded in particular ways is a fundamental task in sensory neuroscience. It is generally believed that a well adapted neural representation can unfold the statistical structure hidden within the stimulus input.

The efficient coding hypothesis was developed following the birth of information theory (Shannon, 1948) and argues that biological sensory systems should maximize the information transfer (Attneave, 1954; Barlow, 1961). This hypothesis has been very seccessful to explain sensory representations (Maddess and Laughlin, 1985; Theunissen and Miller, 1991; Fitzpatrick et al., 1997; Harper and McAlpine, 2004). Experimentally, many studies have also demonstrated that sensory neural codes are indeed adapting to the input distribution statistics for higher coding efficiencies (Brenner et al., 2000; Twer and MacLeod, 2001; Dean et al., 2005; Ozuysal and Baccus, 2012).

There are two key components that need to be clarified before we can formulate an efficient coding problem. The first component is a set of constraint the neural code is facing. Such constraints are often chosen to reflect the natural limitations of a real neuron. From time to time, some constraints may need to be relaxed or removed as a compromise for analytical tractability of the coding problem. The second component is the objective function the neural code is presumably optimizing. Although the mutual information has been the most popular choice, other optimal criteria should not be neglected. In this thesis we aim to extend the efficient coding hypothesis in these two directions and propose a general framework to analytically solve the efficient coding problem.

1.1. Efficient Coding Problem: Constraints

The first key component is the set of constraints that reflects the biological limitations a neuron or the neural population is facing. The simplest and mostly used constraint is a gain control, which limits the maximum output that can be generated by a neuron (Laughlin, 1981; Nadal and Parga, 1994; Brunel and Nadal, 1998; Zhang and Sejnowski, 1999; Pouget et al., 1999; Nikitin et al., 2009; Ganguli and Simoncelli, 2010). A small number of studies have investigated other types of constraint such as the mean firing rate from theoretical perspectives (Nadal and Parga, 1994; Ganguli and Simoncelli, 2014). Despite the low attention received from theoretical studies, the mean output constraint are often used as a regularization term in numerical studies (Olshausen and Field, 1996; Karklin and Simoncelli, 2011; Zhao and Zhaoping, 2011). Another important factor is the neural noise model. The traditional noise model in information theory is the constant Gaussian noise (Laughlin, 1981; Karklin and Simoncelli, 2011; Doi and Lewicki, 2011). In neuroscience however, the canonical model is the Poisson spiking process (Chichilnisky, 2001; Bethge et al., 2002, 2003; Yaeli and Meir, 2010; Ganguli and Simoncelli, 2014). Recently, both sub-Poisson and sup-Poisson spiking behavior are also receiving increasing attention (Churchland et al., 2010; Goris et al., 2014).

1.2. Efficient Coding Problem: Objective Functions

Another component is the utility function that the neural codes is presumably optimized for. A large fraction of previous work assumed that neural representations are tuned to maximize the mutual information they are able to convey about the stimulus values given some overall constraints on available metabolic costs (*e.g.* total number of spikes) (Laughlin, 1981; Seung and Sompolinsky, 1993; Nadal and Parga, 1994; Brunel and Nadal, 1998; Zhang and Sejnowski, 1999; Pouget et al., 1999; Kang et al., 2004; Sharpee et al., 2006; McDonnell and Stocks, 2008; Nikitin et al., 2009; Tkacik et al., 2010; Yarrow et al., 2012; Ganguli and Simoncelli, 2014). This Infomax criterion has been a preferred choice because it does not require making any further assumptions about potential downstream computations and

tasks the encoded stimulus may be involved in. On the other hand, a few studies have taken a downstream perspective and have argued for optimality criteria that consider how well the stimulus information can actually be reconstructed from the neural representations. They often use a metric criterion in terms of the mean squared reconstruction error (Bethge et al., 2002, 2003; Berens et al., 2009; Yaeli and Meir, 2010; Doi and Lewicki, 2011; Ganguli and Simoncelli, 2014). This reconstruction metric has been shown to optimize performance in perceptual estimation and classification tasks (Salinas, 2006). A comparison of these two approaches is summarized in Figure 1. However, a unified comparison and evaluation of these different approaches is currently lacking.

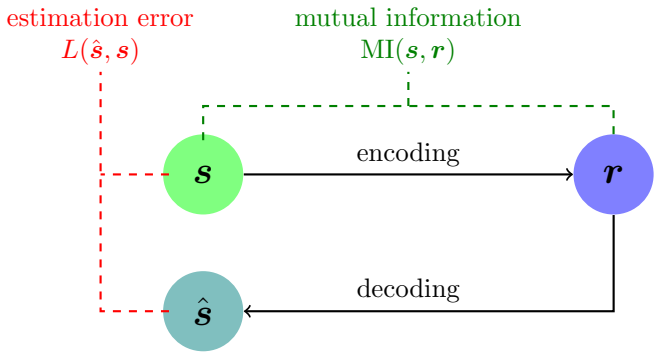

Figure 1: Efficient coding hypothesis: information theoretic approach versus estimation error metric approach.

1.3. Thesis Outline

In this thesis, we aim to have a thoroughly understanding of the optimal neural codes that process natural stimuli and use that to enhance our understanding of the observed neural behaviors. In Chapter 2, we establish the necessary assumptions and present a key statistical tool that plays a central role when evaluating the performance of neural codes and understanding various constraints. These assumptions and the preliminary results will be useful in all of the following chapters.

In Chapter 3, we begin with the traditional efficient coding theory to consider neural codes that maximize mutual information (infomax), yet under a much more general setting. In particular, we propose a framework to derive the infomax neural codes under various com-

bination of biological constraints, such as the distribution of neural noise, the types of constraints on the tuning curves. Important examples of noise models being considered are the Poisson noise model, constant Gaussian noise model. The constraints on the tuning curves include a range constraint which limits the upper and lower bound of neural output and a metabolic cost budget constraint which limits the mean output of a neuron, averaged over the randomness of the input stimulus. Using results from our model, we show that the biologically observed ON-OFF pathway splitting is optimal if a binding metabolic constraint exists.

In Chapter 4, we switch our focus from the constraints on the neural output to the objective functions itself. We provide an initial comparison between different criteria such as the mutual information maximization and decoding error minimization. We introduce a parametric formulation of the efficient coding problem in terms of minimizing the overall reconstruction error according to the L_p norm, as a function of the norm parameter p. We assume reconstruction from a maximum likelihood estimate (MLE) decoder in the asymptotic time limit. Assuming certain noise model, we analytically derive the optimal tuning curve to achieve minimal L_p mean reconstruction error for arbitrary stimulus distributions. This framework includes both the infomax as well as mean-squared error optimal solutions in the limit of $p \to 0$ and $p = 2$ respectively. We first focus on solutions for the optimal tuning curve $h(s)$ of a single (sigmoidal) neuron encoding the stimulus. We then show how the single neuron tuning curve solution can be naturally extended to populations of neurons. Under certain assumptions, the optimal single neuron tuning curve $h(s)$ can be related to an optimal *meta-tuning curve* of the neural population, from which the individual tuning characteristics of the population of neurons can be determined. Using this framework, we investigate the possible underlying principles of various sensory modalities.

In Chapter 5, we further extend our results from Chapter 4 to incorporate multivariate input stimulus. We first generalize the optimal criteria to multivariate case and analytically derive the optimal neural codes for a neural population. Although some additional limitations

will inevitably arise compare to the one dimensional case, our results still offer a good understanding of how are different optimal criteria are related to each other. This result can help us to understand the encoding of pixel values of natural images and the optimal codes for different criteria is compared.

CHAPTER 2 : Models and Methods

In this chapter we present the basic assumptions on the neural encoding and decoding models. Also we will present a few useful statistical tools that will play a key role in evaluating various neural codes.

2.1. Neural Codes: Encoding and Decoding Processes

2.1.1. Encoding Process

We let $\mathbf{s} \in \mathbf{R}^n$ be a n-dimensional stimulus input with prior density $f(\mathbf{s})$. We use $\mathbf{h}(\mathbf{s})$ to represent the neural code which maps the stimulus to m-dimensional output using a population of m neurons. If the input stimulus s is one dimensional, such mapping is a vector $\mathbf{h}(s) = (h_1(s), \ldots, h_m(s))$. If the input stimulus is multivariate, we extend the definition using more variables: $h_k(\mathbf{s}) = g_k \cdot \varphi_k(\mathbf{w}_k^T \mathbf{s})$ where $\mathbf{W}_{n\times m} = (\mathbf{w}_1, \ldots, \mathbf{w}_m)$ is the linear transformation, $\varphi_k(\cdot)$ is the specific activation function and g_k is the gain for the k-th neuron. Together with a certain noise model (see Section 2.1.2), this completes the Linear-Nonlinear-Noise model.

In particular, we will consider the simplest scenario of encoding a one dimensional stimulus ($n = 1$) using a single neuron ($m = 1$) or multiple neurons ($m > 1$). In these two cases, the linear projection $\mathbf{W}$ is scaler w and the gain multiplier g_k can be simultaneously incorporated using a simpler notation $h_k(s) = g_k \cdot \varphi_k(w \cdot s)$. In the most generic scenario, we consider the harder problem of encoding a high dimensional stimulus ($n > 1$). To avoid decoding ambiguity issues, we assume the neural population is complete ($m = n$) or over-complete ($m > n$). In Figure 2 we compare between these three described cases.

2.1.2. Neural Noise Models

When generating their output, neurons in the brain are known to be noisy. The actual output $\mathbf{r}$ is stochastic even if the same stimulus $\mathbf{s}$ is presented. Such stochasticity is determined

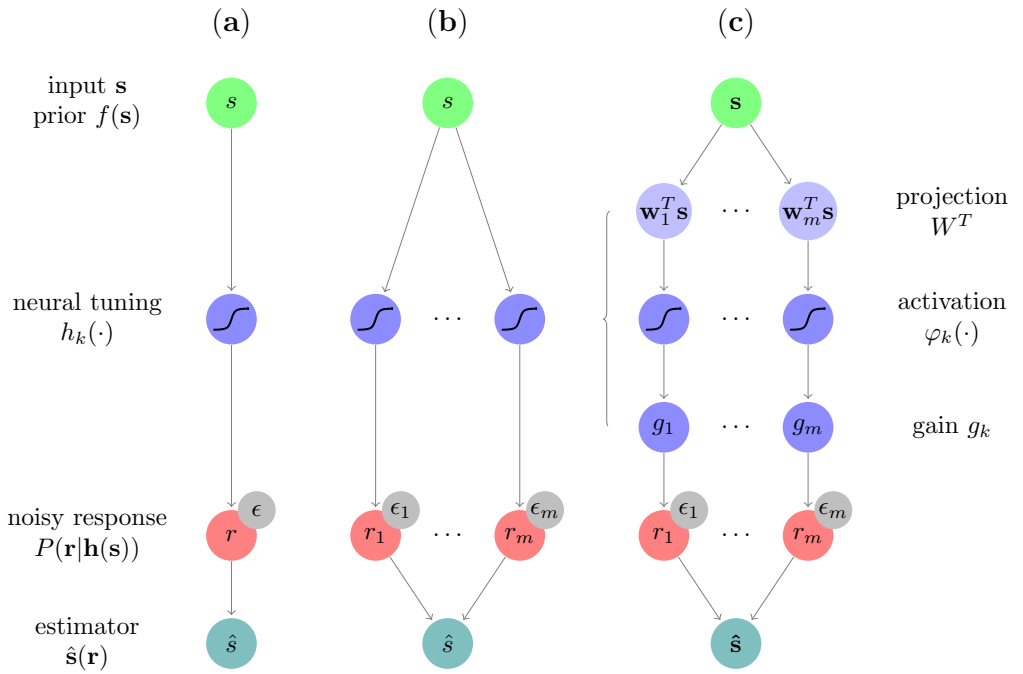

Figure 2: The encoding process of a Linear-Nonlinear-Noise model. We show (a) the simplest case where a single neuron is encoding a one-dimensional stimulus (b) a neural population is encoding a one-dimensional stimulus and the generic case (c) a complete or over complete neural population is encoding a multivariate stimulus.

by a probabilistic model $p(\mathbf{r}|\mathbf{h}(\mathbf{s}))$. There are a couple of assumptions that need to be made about this probabilistic model. First of all, we assume the response is centralized around the desired output $\mathbf{h}(\mathbf{s})$ and the mean is equal to the desired output $\langle\mathbf{r}\rangle = \mathbf{h}(\mathbf{s})$. Second, we assume each dimension of the output is independent from each other, i.e.

$$p(\mathbf{r}|\mathbf{h}(\mathbf{s})) = \prod_{i=1}^{m} p(r_i|h_i(\mathbf{s})) \tag{2.1}$$

In particular, different noise models have been proposed based on how the neural output is defined. For example, the spike timing is often modeled as a Poisson process. As another example, the membrane potential of a neuron is subject to noises from many independent

sources. This gives us two simple noise models to begin with

$$\textbf{Poisson (P):} \quad N_i \sim \text{Poisson}(h(s)T) \quad \text{and} \quad r_i = N_i/T \tag{2.2}$$

$$\textbf{constant Gaussian (cG):} \quad r_i \sim \text{Normal}(h(s), V_0/T), \tag{2.3}$$

where T is the length of the time window for encoding. The first case corresponds to a typical Poisson spiking model and complete the trio of the canonical Linear-Nonlinear-Poisson cascade model (Chichilnisky, 2001). Over time T, the total number of spikes N_i elicited from a neuron should follow a Poisson distribution and it is easy to verify that $\langle N_i \rangle = \mathbf{Var}[N_i] = h(s)T$ which leads to $\langle r_i \rangle = h(s)$ and $\mathbf{Var}[r_i] = h(s)/T$. Compare this with the constant Gaussian noise case, we consider a more generalized noise model parameterized by α

$$\textbf{general Gaussian (gG):} \quad r_i \sim \text{Normal}(h(s), V_0 \cdot h(s)^\alpha/T), \tag{2.4}$$

With special choices of α, we retrieve good approximations of the constant Gaussian model (cG, $\alpha = 0$) or Poisson noise model (P, $\alpha = 1$) respectively.

2.1.3. Decoding Process

Although it is not necessary to choose a decoder when using information theoretic metrics, it is crucial to choose a proper decoder so that we can evaluate the performance of neural codes defined by estimation error. For most part of the thesis, we assume the maximum likelihood estimator (MLE) $\hat{s}(r)$, maximizes the likelihood

$$\hat{\mathbf{s}}_{\text{MLE}}(\mathbf{r}) = \arg\max_{\mathbf{s}} p(\mathbf{r}|\mathbf{s}) \tag{2.5}$$

The MLE has nice statistical properties *e.g.* asymptotically unbiased and efficient. More discussion can be found in Section 2.2. Another competitive decoder is the *Maximum A*

Posteriori Estimator (MAPE), which also includes the prior distribution into the picture

$$\hat{\mathbf{s}}_{\text{MAPE}}(\mathbf{r}) = \arg\max_{\mathbf{s}} p(\mathbf{r}|\mathbf{s}) f(\mathbf{s}) \tag{2.6}$$

In the long time asymptotic, the optimal decoder naturally converges to the maximum likelihood estimator because with sufficient evidence accumulation, the neural signal will be much more reliable than the prior information. On the other hand, with short encoding time, it is often the case that a Bayesian (and usually biased) decoder will perform better (Wei and Stocker, 2015). Unfortunately it is very hard to analytically optimize the short term neural code and we often need to rely on numerical tools and methods (Bethge et al., 2003; Nikitin et al., 2009). In addition, the derivation of the optimal Bayesian decoder can be intractable for arbitrary prior. For these reasons we will focus on using the MLE as an ideal decoder to complete the encoding-decoding pipeline.

2.2. Fisher Information and Neural Codes

The concept of Fisher Information provides a statistical characterization of how well a random variable $\mathbf{r}$ can be used to estimate an underlying parameter s under a stochastic model $p(\mathbf{r}|\mathbf{s})$. For the purpose of generality, we will present all definitions and properties in multivariate form and we assume $\mathbf{s} \in \mathbf{R}^n$ and $\mathbf{r} \in \mathbf{R}^m$. Some important applications including the one dimensional stimulus scenario is just a simple case of the general definition.

In statistics, the *score function* is defined as the $n \times 1$ gradient vector of the log-likelihood

$$\boldsymbol{\theta}(\mathbf{s}, \mathbf{r}) = \nabla_{\mathbf{s}} \log p(\mathbf{r}|\mathbf{s}) = \left(\frac{\partial}{\partial s_1} \log p(\mathbf{r}|\mathbf{s}), \ldots, \frac{\partial}{\partial s_n} \log p(\mathbf{r}|\mathbf{s}) \right)^T \tag{2.7}$$

Please note that we will omit the subscripting variable $\mathbf{s}$ in $\nabla_{\mathbf{s}}$ when obvious in the rest of the thesis. Now the Fisher Information matrix has size $n \times n$ and is defined as (see Cover

and Thomas (1991))

$$\mathcal{I}(\mathbf{s}) = \left\langle \boldsymbol{\theta}(\mathbf{s},\mathbf{r}) \cdot \boldsymbol{\theta}(\mathbf{s},\mathbf{r})^T \middle| \mathbf{s} \right\rangle_{p(\mathbf{r}|\mathbf{s})} \tag{2.8}$$

The Fisher Information Matrix plays an key role in relating the neural codes and the objective functions – from information theoretic quantities to error metrics. This will be elaborated below.

2.2.1. Population of Neurons with Independent Noise

First we study how does each neuron in the population contribute to the Fisher Information matrix. In our model, each neuron has an output r_i that occupies one dimension of the output vector $\mathbf{r}$. Throughout this thesis we will mainly study the case when each neuron has independent noise. In this case,

$$p(\mathbf{r}|\mathbf{s}) = \prod_{k=1}^{m} p(r_k|\mathbf{s}) \quad \Rightarrow \quad \boldsymbol{\theta}(\mathbf{s},\mathbf{r}) = \nabla_{\mathbf{s}} \left(\sum_k \log p(r_k|\mathbf{s}) \right) = \sum_k \boldsymbol{\theta}(\mathbf{s}, r_k) \tag{2.9}$$

To show this, we using the definition of Fisher Information Matrix

$$\mathcal{I}_{\text{total}}(\mathbf{s}) = \left\langle \boldsymbol{\theta}(\mathbf{s},\mathbf{r}) \cdot \boldsymbol{\theta}(\mathbf{s},\mathbf{r})^T \middle| \mathbf{s} \right\rangle = \left\langle \left(\sum_k \boldsymbol{\theta}(\mathbf{s}, r_k) \right) \cdot \left(\sum_l \boldsymbol{\theta}(\mathbf{s}, r_l) \right)^T \middle| s \right\rangle \tag{2.10}$$

Due to the independence between r_k and r_l for $k \neq l$ when conditioned on any given s,

$$\left\langle \boldsymbol{\theta}(\mathbf{s}, r_k) \cdot \boldsymbol{\theta}(\mathbf{s}, r_l)^T \middle| s \right\rangle = \langle \boldsymbol{\theta}(\mathbf{s}, r_k) | s \rangle \cdot \langle \boldsymbol{\theta}(\mathbf{s}, r_l) | s \rangle^T = \mathbf{0} \tag{2.11}$$

where the second inequality is because each index $\theta_i(\mathbf{s}, r_k)$ for any i, k is

$$\left\langle \frac{\partial}{\partial s_i} \log p(r_k|\mathbf{s}) \middle| \mathbf{s} \right\rangle = \int \frac{\frac{\partial}{\partial s_i} p(r_k|\mathbf{s})}{p(r_k|\mathbf{s})} \cdot p(r_k|\mathbf{s}) dr_k = \frac{\partial}{\partial s_i} \left(\int p(r_k|\mathbf{s})\, dr_k \right) = 0. \tag{2.12}$$

As a conclusion, the total Fisher Information for a population of neurons with independent Poisson/constant Gaussian noise is equal to the linear sum of the Fisher information of each

neuron:

$$\mathcal{I}_{\text{total}}(\mathbf{s}) = \sum_{k=1}^{m} \left\langle \boldsymbol{\theta}(\mathbf{s}, r_k) \cdot \boldsymbol{\theta}(\mathbf{s}, r_k)^T \middle| s \right\rangle = \sum_{k=1}^{m} \mathcal{I}_k(\mathbf{s}) \tag{2.13}$$

This conclusion allow us to calculate the Fisher information of a neural population by simply calculating the Fisher information for each neuron one at a time. With this benefit, we analyze how does the Fisher information depend on the neural noise model for a single neuron

2.2.2. Fisher Information for Poisson Noise Model

The first model is the Poisson spiking model. If the neuron elicits a random number of spikes r during a given time window T is a Poisson random variable with rate $T \cdot h(s)$

$$P(r = N|\mathbf{s}) = \frac{1}{N!} \left(T \cdot h(\mathbf{s})\right)^N \exp(-T \cdot h(\mathbf{s})) \tag{2.14}$$

$$\log P(r = N|\mathbf{s}) = -\log(N!) + N \log\left(T \cdot h(\mathbf{s})\right) - T \cdot h(\mathbf{s}) \tag{2.15}$$

$$\nabla_{\mathbf{s}} \log P(r = N|\mathbf{s}) = \nabla h(\mathbf{s}) \left(\frac{N}{h(\mathbf{s})} - T\right) \tag{2.16}$$

For Poisson random variable N with rate $T \cdot h(s)$ we use some simple facts to calculate the Fisher information matrix

$$\langle N \rangle = T \cdot h(\mathbf{s}), \quad \left\langle N^2 \right\rangle = T \cdot h(\mathbf{s}) + \left(T \cdot h(\mathbf{s})\right)^2 \tag{2.17}$$

$$\mathcal{I}(\mathbf{s}) = \nabla h(\mathbf{s}) \nabla h(\mathbf{s})^T \left\langle \left(\frac{N}{h(\mathbf{s})} - T\right)^2 \right\rangle = T \cdot \frac{\nabla h(\mathbf{s}) \nabla h(\mathbf{s})^T}{h(\mathbf{s})} \tag{2.18}$$

2.2.3. Fisher Information for Constant Gaussian Noise Model

In the second model, we relax the neural output range from non-negative integers to all real values. In this general Gaussian noise case, the additive noise in each unit time window is $V_0 h(\mathbf{s})^\alpha$ therefore the output r over a time window of length T is a Gaussian random

variable with mean $\mu(\mathbf{s}) = T \cdot h(s)$ and variance $V(\mathbf{s}) = T \cdot V_0 h(\mathbf{s})^\alpha$

$$p(r|\mathbf{s}) = \frac{1}{\sqrt{2\pi V(\mathbf{s})}} \exp\left(-\frac{1}{2V(s)}(r-\mu(\mathbf{s}))^2\right) \tag{2.19}$$

$$\log p(r|\mathbf{s}) = -\frac{1}{2}\log(2\pi V(\mathbf{s})) - \frac{1}{2V(\mathbf{s})}(r-\mu(\mathbf{s}))^2 \tag{2.20}$$

$$\nabla_{\mathbf{s}} \log p(r|\mathbf{s}) = -\frac{1}{2}\frac{\nabla V(\mathbf{s})}{V(\mathbf{s})} + \frac{\nabla V(\mathbf{s})}{2V(\mathbf{s})^2}(r-\mu(\mathbf{s}))^2 + \frac{\nabla\mu(\mathbf{s})}{V(\mathbf{s})}(r-\mu(\mathbf{s})) \tag{2.21}$$

Use some simple fact about the random variable r, we can calculate the Fisher information for a neuron with Gaussian noise

$$\langle r - h(\mathbf{s})\rangle = 0, \quad \left\langle (r-h(\mathbf{s}))^2 \right\rangle = V(\mathbf{s}), \tag{2.22}$$

$$\left\langle (r-h(\mathbf{s}))^3 \right\rangle = 0, \quad \left\langle (r-h(\mathbf{s}))^4 \right\rangle = 3V(\mathbf{s})^2 \tag{2.23}$$

$$\mathcal{I}(s) = \frac{\nabla\mu(\mathbf{s})\nabla\mu(\mathbf{s})^T}{V(\mathbf{s})} + \frac{1}{2}\frac{\nabla V(\mathbf{s})\nabla V(\mathbf{s})^T}{V(\mathbf{s})^2} = T \cdot \frac{\nabla h(\mathbf{s})\nabla h(\mathbf{s})}{V_0 h(\mathbf{s})^\alpha} + O(1) \tag{2.24}$$

In Eq. (2.18) and Eq. (2.24) we have derived the Fisher Information of a single neuron with Poisson or constant Gaussian noise model. Generally speaking, each neuron contributes a positive semidefinite matrix of rank one towards the total Fisher information matrix. In order to have a non-degenerate Fisher information matrix (i.e. with rank n), it is necessary for the neural population to be complete $m = n$ or over complete $m > n$.

2.2.4. Fisher Information for Linear-Nonlinear Model

If we also incorporate the linear transformation phase of encoding, the tuning curve becomes $h(\mathbf{s}) = g \cdot \varphi(\mathbf{w}^T\mathbf{s})$ where g controls the gain of the tuning curve and φ is a sigmoidal function bounded between $[0, 1]$. If we use the Poisson noise model, the Fisher information of such neuron is

$$\nabla h(\mathbf{s}) = g \cdot \varphi'(\mathbf{w}^T\mathbf{s})\mathbf{w} \quad \Rightarrow \quad \mathcal{I}(\mathbf{s}) = T \cdot g \cdot \frac{\varphi'(\mathbf{w}^T\mathbf{s})^2}{\varphi(\mathbf{w}^T\mathbf{s})}\mathbf{w}\mathbf{w}^T \tag{2.25}$$

If we use the generalized Gaussian noise model, the Fisher information becomes

$$\nabla h(\mathbf{s}) = g \cdot \varphi'(\mathbf{w}^T\mathbf{s})\mathbf{w} \quad \Rightarrow \quad \mathcal{I}(\mathbf{s}) = T \cdot \frac{g^{2-\alpha}}{V_0} \frac{\varphi'(\mathbf{w}^T\mathbf{s})^2}{\varphi(\mathbf{w}^T\mathbf{s})^\alpha} \mathbf{w}\mathbf{w}^T + O(1) \tag{2.26}$$

For general $\alpha < 2$, the value of Fisher information can be further simplified

$$\varphi(t) = \tilde{\varphi}(t)^{2/(2-\alpha)} \quad \Rightarrow \quad \mathcal{I}(\mathbf{s}) \propto T \cdot g^{2-\alpha} \cdot \tilde{\varphi}'(\mathbf{w}^T\mathbf{s})^2 \mathbf{w}\mathbf{w}^T + O(1) \tag{2.27}$$

Such trick is known as the Variance Stablization Transformation (Cover and Thomas, 1991). The transformed functions $\tilde{\varphi}$ are still a sigmoidal function with saturation range $0 \leq \tilde{\varphi} \leq 1$, yet the Fisher information has a much simpler form. Here we remark that the new form of Fisher information rate is the same as the constant Gaussian noise case ($\alpha = 0$) except the part that depends on the neural gain $g^{2-\alpha}$.

2.3. Fisher Information and Objective Functions

2.3.1. Mutual Information

One possible measurement of neural coding quality is the mutual information. Let us assume random variables $\mathbf{s}, \mathbf{r}$ has density $f(\mathbf{r}, \mathbf{s})$ and marginal distributions $f(\mathbf{r}), f(\mathbf{s})$ respectively. Mutual information is a concept of information theory which measures the mutual dependence between two random variable $\mathbf{r}, \mathbf{s}$.

$$\mathrm{MI}(\mathbf{r}, \mathbf{s}) = \iint f(\mathbf{r}, \mathbf{s}) \log \frac{f(\mathbf{s}, \mathbf{r})}{f(\mathbf{r}) f(\mathbf{s})} \, d\mathbf{r} d\mathbf{s} \tag{2.28}$$

When evaluating neural codes, the biggest advantage to use mutual information is because it does not require any assumptions on how the neural representation $\mathbf{r}$ is used in downstream tasks. The link between mutual information $\mathrm{MI}(\mathbf{r}, \mathbf{s})$ and the Fisher information matrix was established by Brunel and Nadal (1998).

$$\mathrm{MI}(\mathbf{r}, \mathbf{s}) = \frac{1}{2} \left\langle \log \det \mathcal{I}(\mathbf{s}) \right\rangle_{\mathbf{s}} + const. \tag{2.29}$$

Here we will not repeat the careful and delicate derivation in their results but the main idea is based on the fact that an efficient and unbiased estimator $\hat{\mathbf{s}}$ is approximately distributed as a Gaussian with mean $\mathbf{s}$ and covariance $\mathcal{I}(\mathbf{s})^{-1}$. The conditional entropy of such Gaussian random variable is locally $1/2 \cdot \log\det(\mathcal{I}(\mathbf{s})^{-1}) + const$ and by averaging the local conditional entropy, we can get the mutual information. The mutual information objective (infomax criterion) can be achieved by maximizing the right side of Eq. (2.29). For a more complete work regarding the relationship between Fisher information and the mutual information, the reader is referred to Wei and Stocker (2016)

2.3.2. Cramer-Rao Lower Bound

Another possible way to measure neural coding quality is to use the L_2 norm of the error vector $\hat{\mathbf{s}} - \mathbf{s}$. Such L_2 norm is related to the Fisher information matrix via the Cramer-Rao lower bound (Cover and Thomas, 1991). For any unbiased estimator $\hat{\mathbf{s}}(\mathbf{r})$, e.g. the maximum likelihood estimator (MLE),

$$\mathbf{cov}[\hat{\mathbf{s}}(\mathbf{r}) - \mathbf{s} \,|\, \mathbf{s}] \geq_{\mathrm{M}} \mathcal{I}(\mathbf{s})^{-1} \tag{2.30}$$

where the matrix inequality $\geq_{\mathrm{M}}$ is defined in the sense that $\mathbf{cov}[\hat{\mathbf{s}} - \mathbf{s} \,|\, \mathbf{s}] - \mathcal{I}(\mathbf{s})^{-1}$ is positive semidefinite. As a lower bound, the Cramer-Rao bound can be attained by the MLE $\hat{\mathbf{s}}(\mathbf{r})$ due to is asymptotic efficiency (Cover and Thomas, 1991).

In order to calculate the mean L_2 error, one can find the attainable lower bound both locally at a given point $\mathbf{s}$ or globally averaged over all $\mathbf{s}$, by taking the trace of the covariance matrix

$$\langle \|\hat{\mathbf{s}} - \mathbf{s}\|^2 \,|\, \mathbf{s} \rangle_{\mathbf{r}} = \mathrm{tr}\left[\mathbf{cov}[\hat{\mathbf{s}}(\mathbf{r}) - \mathbf{s} \,|\, \mathbf{s}]\right] \geq \mathrm{tr}\left[\mathcal{I}(\mathbf{s})^{-1}\right]. \tag{2.31}$$

$$\left\langle \|\hat{\mathbf{s}} - \mathbf{s}\|^2 \right\rangle_{\mathbf{r},\mathbf{s}} \geq \left\langle \mathrm{tr}\left[\mathcal{I}(\mathbf{s})^{-1}\right] \right\rangle_{\mathbf{s}} \tag{2.32}$$

Compare this with Eq. (2.29), we now derive another way of evaluating the Fisher information matrix. In order to minimize the mean L_2 error, one should minimize the right

side of Eq. (2.32). For a more complete work regarding the relationship between Fisher information and the Cramer-Rao lower bound, the reader is referred to Pilarski and Pokora (2015)

2.3.3. Asymptotic L_p Limit

To evaluate the decoding error $\hat{\mathbf{s}} - \mathbf{s}$, a natural generalization of L_2 norm is the L_p norm (or semi-norm when $p < 1$) for other values of p. However, a direct generalization to the multivariate case would fail because the L_p norm is not rotational invariant unless $p = 2$. In other words, the L_p norm of $\hat{\mathbf{s}} - \mathbf{s}$ depends on the choice of coordinate system and makes it impossible to fairly compare different neural codes.

To avoid this problem, we use a different definition of L_p error in a rotational invariant way. For the random variable $\hat{\mathbf{s}} - \mathbf{s}$, asymptotically distributed as Gaussian with mean $\mathbf{0}$ and variance $\mathcal{I}(\mathbf{s})^{-1}$. We denote the eigenvalue of $\mathcal{I}(\mathbf{s})^{-1}$ is $\lambda_1, \ldots, \lambda_n$, then we define

$$\|\mathcal{I}(\mathbf{s})^{-1}\|_p = \sum_i \lambda_i^{p/2} \tag{2.33}$$

In one dimension, this automatically falls back to the ordinary L_p loss measurement

$$\|\mathcal{I}(s)^{-1}\|_p = \mathcal{I}(s)^{-p/2} \tag{2.34}$$

Because the eigenvalues of the covariance matrix are invariant under rotations, this is indeed a unitary invariance choice of L_p metric in high dimensional space. When $p = 2$, we can retrieve

$$\|\mathcal{I}(\mathbf{s})^{-1}\|_p = \sum_i \lambda_i = \mathrm{tr}\left[\mathcal{I}(\mathbf{s})^{-1}\right] \tag{2.35}$$

which is identical to the Cramer-Rao lower bound discussed above. In order to obtain the optimal L_p population code, one can minimize the mean L_p norm of the uncertainty

covariance $\mathcal{I}(\mathbf{s})^{-1}$

$$\langle \|\hat{\mathbf{s}}(\mathbf{r}) - \mathbf{s}\|^p \rangle_{\mathbf{r},\mathbf{s}} \approx const(p) \cdot \left\langle \|\mathcal{I}(\mathbf{s})^{-1}\|_p \right\rangle \tag{2.36}$$

This family of optimization problems with various value of p can provide a natural connection between two traditional optimal criteria – the infomax and MMSE (L_2-min). In the limit of $p \to 0$, we can use the replica trick to show that minimizing the right side of Eq. (2.36) is equivalent to maximizing the mutual information term in Eq. (2.29).

$$\lim_{p\to 0} \frac{\sum_i \lambda_i^{p/2} - 1}{p} = \frac{1}{2} \sum_i \log \lambda_i = -\frac{1}{2} \log \det \mathcal{I}(\mathbf{s}) \tag{2.37}$$

CHAPTER 3 : Infomax Codes Under Energy Constraints

3.1. Introduction

The efficient coding hypothesis (Attneave, 1954; Barlow, 1961) plays a fundamental role in understanding neural codes, particularly in early sensory processing. Going beyond the original idea of redundancy reduction by Barlow (1961), efficient coding has become a general conceptual framework for studying optimal neural coding (Linsker, 1988; Atick. and Redlich, 1990; Atick, 1992; Rieke et al., 1995; Olshausen and Field, 1996; Bell and Sejnowski, 1997; Simoncelli and Olshausen, 2001; Gottschalk, 2002; Harper and McAlpine, 2004; McDonnell and Stocks, 2008; Karklin and Simoncelli, 2011; Wei and Stocker, 2016). Efficient coding hypothesizes that the neural code is such that it maximizes the information conveyed about the stimulus variable. Any formulation of efficient coding necessarily relies on a set of constraints. These constraints can come in various ways as reflected by the many real world limitations neural systems are facing. For examples, noise, limited metabolic energy budgets, constraints on the shape of tuning curves, the number of neurons in the system *etc.* all limit the dynamic range and accuracy of the neural code.

Previous studies mainly considered only a small subset of these constraints. For example, the original proposal of redundancy reduction by Barlow focused on utilizing the dynamical range of the neurons efficiently (Barlow, 1961, 2001), but did not address the problem of noise and energy consumption. Some studies explicitly dealt with the metabolic costs of the system but did not consider the constraints imposed by the limited firing rates of neurons as well as their detailed tuning properties (Levy and Baxter, 1996; Olshausen and Field, 1996; Laughlin et al., 1998; Balasubramanian et al., 2001). Histogram equalization has been proposed as the mechanism in determining the optimal tuning curve of a single neuron with monotonic response characteristics (Laughlin, 1981). However, this result relies on restrictive assumptions of the neural noise and does not take metabolic costs into consideration. In terms of neural population coding, most previous studies have focused

on bell-shaped tuning curves. Optimal neural coding for monotonic tuning curves have received only little attention (Ganguli and Simoncelli, 2014; Kastner et al., 2015).

We developed a formulation of efficient coding that explicitly deals with multiple biologically relevant constraints, including neural noise, limited range of the neural output, and metabolic constraints. We use our formulation to study neural codes based on monotonic response characteristics that have been frequently observed in biological neural systems. We were able to derive analytical solutions for a wide range of conditions in the small noise limit. We present results for neural pools of different sizes, including the cases of a single neuron, pairs of neurons, and larger neural populations. The results are in general agreements with observed coding schemes for monotonic tuning curves. The results also provides various quantitative predictions which are readily testable with targeted physiology experiments.

3.2. Model Assumptions

In this chapter, we start with the simple case where a scaler stimulus s with prior $f(s)$ is encoded by a single neuron. To model the neural response for stimulus s, we denote the mean output level as $h(s)$. As we have discussed (see Section 2.1.1), such value $h(s)$ is a deterministic mapping from s and could be the mean firing rate in the context of rate coding or the just the mean membrane potential. The actual response r is noisy and follows one of the possible noise models (see Section 2.1.2). Throughout this chapter, we constrain ourself to neural codes with monotonic response functions.

We formulate the efficient coding problem as the neural code seeks to maximize the mutual information between the stimulus and the response, *e.g.*, $\mathrm{MI}(s, r)$ (Linsker, 1988). To complete the formulation of this problem, it is crucial to choose a set of constraints which characterizes the limited resource available to the neural system. One constraint is the finite range of the neural output (Laughlin, 1981). Another plausible constraint is on the mean metabolic cost (Levy and Baxter, 1996; Olshausen and Field, 1996; Laughlin et al., 1998; Balasubramanian et al., 2001), which limits the mean activity level of neural output. Under

these constraints, the efficient coding problem is mathematically formulated as following:

$$\begin{aligned} \text{maximize} \quad & \mathrm{MI}(s, r) \\ \text{subject to} \quad & 0 \le h(s) \le r_{\max}, \quad h'(s) \ge 0 && \text{(range constraint)} \\ & \mathbf{E}_s[K(h(s))] \le K_{\text{total}} && \text{(metabolic constraint)} \end{aligned}$$

We seek the optimal response function $h(s)$ under various choices of the neural noise model $P(r|h(s))$ and certain metabolic cost function $K(h(s))$, as discussed below.

Neural Noise Models: Neural noise in early sensory area can often be well characterized by a Poisson distribution (Tomko and Crapper, 1974; Tolhurst et al., 1981). Under the Poisson noise model, the number of spikes N_T over a duration of T is a Poisson random variable with mean $h(s)T$ and variance $h(s)T$. In the long T limit, the mean response $r = N_T/T$ approximately follows a Gaussian distribution

$$r \sim \mathcal{N}(h(s), h(s)/T) \tag{3.1}$$

Non-Poisson noise have also been observed where the variance of response N_T can be greater or smaller than the mean firing rate (Tomko and Crapper, 1974; Tolhurst et al., 1981; Churchland et al., 2010; Goris et al., 2014). Therefore we consider a more generalized family of noise models parametrized by α

$$r \sim \mathcal{N}(h(s), h(s)^{\alpha}/T) \tag{3.2}$$

This generalized family of noise model naturally includes the additive Gaussian noise case (when $\alpha = 0$), which is useful to describe the stochasticity of the membrane potential.

Metabolic Cost: We consider the metabolic cost K as a function of the neural output

$$K(h(s)) = h(s)^{\beta} \tag{3.3}$$

where $\beta > 0$ is a parameter to model how does the energy cost scale up as the neural output is increasing. For a single neuron we will demonstrate with the general energy cost function but when we generalize to the case of multiple neurons, we will use a linear model suggested by Attwell and Laughlin (2001) for clarity

$$K(h(s)) = K_0 + K_1 h(s) \tag{3.4}$$

In the context of rate coding, $K_0 = K(0)$ can be understood as the energy cost per unit time to maintain a resting neuron and K_1 is the energy cost for each extra spike per unit time. Because the metabolic constraint is also linear in $K(h(s))$, this is equivalent to the above cost function with $\beta = 1$ and properly adjusted K_{total}.

3.3. Optimal Code for a Single Neuron

3.3.1. Derivation of the Optimal $h_*(s)$

This optimization problem can be greatly simplified thanks to the fact that it is invariant with respect to a re-parameterization of the stimulus variable $u = F(s)$ for any invertible transformation F. First of all, we can prove that the mutual information is indeed invariant under parameter transformation by applying the information processing inequality twice together with the fact that $F(s)$ is invertible:

$$I(s, r) \geq I(F(s), r) \geq I(F^{-1}(F(s)), r) = I(s, r). \tag{3.5}$$

Second, we can show that the energy constraint is also invariant:

$$g(u) \stackrel{\text{def}}{=} h(F^{-1}(u)) = h(s), \quad f(u)\, du = f(s)\, ds \tag{3.6}$$

$$\mathbf{E}_u[K(g(u))] = \int_0^1 K(g(u)) f(u)\, du = \int_{-\infty}^{\infty} K(h(s)) f(s)\, ds = \mathbf{E}_s[K(h(s))] \tag{3.7}$$

To take the most advantage out of this, we choose such transformation to be $F(s) = \int_{-\infty}^{s} f(t)\, dt$ to be the cumulative distribution of the prior $f(s)$. In this way, the transformed

variable $u = F(s)$ follows the uniform distribution $U \sim \mathcal{U}[0,1]$. Now it suffices to solve the following new problem which optimizes $g(u)$ for the uniformly distributed input u. Once the optimal form of $g_*(u)$ is obtained, the optimal $h_*(s)$ is naturally given by $g_*(F(s))$

$$\begin{aligned} \text{maximize} \quad & \mathrm{MI}(u, r) \\ \text{subject to} \quad & 0 \le g(u) \le r_{\max}, \quad g'(u) \ge 0 \\ & \mathbf{E}_u[K(g(u))] \le K_{\text{total}} \end{aligned}$$

To solve this simplified problem, first we express the objective function in terms of $g(u)$. In the small noise limit (large T), the Fisher information $\mathcal{I}(u)$ for a neuron with generalized Gaussian noise is calculated (see Chapter 2)

$$\mathcal{I}(u) = \frac{T}{V_0} \frac{g'(u)^2}{g(u)^\alpha} + O(1) \tag{3.8}$$

$$\mathrm{MI}(u, r) = H(U) + \frac{1}{2} \int f(u) \log \mathcal{I}(u)\, du = \frac{1}{2} \int_0^1 \log \frac{g'(u)^2}{g(u)^\alpha}\, du + \log(T/V_0) + O(1/T) \tag{3.9}$$

where $H(U) = 0$ is the entropy and $f(u) = 1_{\{0 \le u \le 1\}}$ is the density of the uniform distribution. Furthermore, each constraints can be rewritten as integrals of $g'(u)$ and $g(u)$ respectively:

$$g(1) - g(0) = \int_0^1 g'(u)\, du \le r_{\max} \tag{3.10}$$

$$\mathbf{E}_u[K(g(u))] = \int_0^1 g(u)^\beta\, du \le K_{\text{total}} \tag{3.11}$$

By throwing away irrelevant constant terms and use Lagrangian multiplier method, the

optimization problem is now simplified to

$$\text{maximize} \quad \int_0^1 L(g(u), g'(u))\, du \tag{3.12}$$

$$\text{where} \quad L(g(u), g'(u)) = \frac{1}{2}\log\frac{g'(u)^2}{g(u)^\alpha} - \lambda_1 g'(u) - \lambda_2 g(u)^\beta \tag{3.13}$$

This problem can be analytically solved by using the Euler-Lagrange equation. In particular, because the Lagrangian $L(g(u), g'(u))$ does not have explicit u dependency, we can apply Beltrami's identity, which is a special form of the Euler-Lagrange equation

$$const = L - g' \cdot \frac{\partial L}{\partial g'} = \left[\frac{1}{2}\log\frac{g'(u)^2}{g(u)^\alpha} - \lambda_1 g'(u) - \lambda_2 g(u)^\beta\right] - g'(u)\cdot\left[\frac{1}{g'(u)} - \lambda_1\right] \tag{3.14}$$

$$\Rightarrow \quad const = \frac{1}{2}\log\frac{g'(u)^2}{g(u)^\alpha} - \lambda_2 g(u)^\beta \tag{3.15}$$

We substite $g(u) = \tilde{g}(u)^{1/\beta}$ and derive an ordinary differential equation (ODE) on $\tilde{g}(u)$ which can be easily solved by separating variables.

$$const = \log\left(\frac{d\tilde{g}}{du}\cdot\tilde{g}^{q-1}\right) - \tilde{\lambda}\tilde{g}(u) \quad \Rightarrow \quad C\, du = \tilde{g}(u)^{q-1}\exp\left(-\tilde{g}\right)\, d\tilde{g} \tag{3.16}$$

where $q = (1-\alpha/2)/\beta$. The solution must take the form of

$$\tilde{g}_*(u) = \frac{1}{a}\gamma_q^{-1}\left(u\gamma_q(b)\right), \quad g_*(u) = \left[\frac{1}{a}\gamma_q^{-1}\left(u\gamma_q(b)\right)\right]^{1/\beta}, \quad h_*(s) = g_*(F(s)) \tag{3.17}$$

$$\text{where } \gamma_q(x) = \int_0^x z^{q-1}\exp(-z)\, dz. \tag{3.18}$$

The function $\gamma_q(x)$ is called the incomplete gamma function of parameter q and γ_q^{-1} is its inverse function. The constants a, b can be determined by satisfying both the sigmoidal and metabolic constraints. Next we present the more intuitive conclusions about a, b while leaving the detailed proof in the appendix (see Section 3.7).

3.3.2. Interpretation of the Optimal $h_*(s)$

Due to the relative difference of $r_{\max}$ and K_{total}, the sigmoidal constraint and the metabolic constraint can be either binding or non-binding. Depending on the relative strength of each constraint:

- **Range constraint dominates:** This is the case when there is more than sufficient energy to achieve the optimal solution $K_{\text{total}} \geq K_{\text{thre}}$. There is a threshold value K_{thre} beyond which the metabolic constraint will become non-binding. The exact value of K_{thre} depends on the model parameters $r_{\max}$, α and β. As a loose estimation, if $K_{\text{total}} \geq r_{\max}^{\beta}$, the metabolic constraint is automatically satisfied for any α. In this case:

$$b \to 0^+, \quad a = b/r_{\max}^{\beta}, \quad g(u) = r_{\max} \cdot u^{1/q} \tag{3.19}$$

  This is because when b is small (so $x = u\gamma_q(b)$ is also small), we have an good approximation of

$$\gamma_q(x) \approx \int_0^x z^{q-1}\, dz = \frac{1}{q}x^q, \quad \gamma_q^{-1}(y) \approx q^{1/q}y^{1/q} \tag{3.20}$$

- **Both constraints:** This is the general case when K_{total} is about the same magnitude as $r_{\max}$. We choose $a = b/r_{\max}^{\beta}$ to satisfy the range constraint and b is set to the minimum value for which the metabolic constraint is satisfied.

- **Metabolic constraint dominates:** This happens when $K_{\text{total}} \ll r_{\max}^{\beta}$. In this case we choose $a = b/r_{\max}^{\beta}$ and b is often very large.

3.3.3. Properties of the Optimal $h(s)$

We have predicted the optimal response function for arbitrary values of α (which constrain the noise) and β (which quantifies the cost). Here we specifically focus on a few biologically

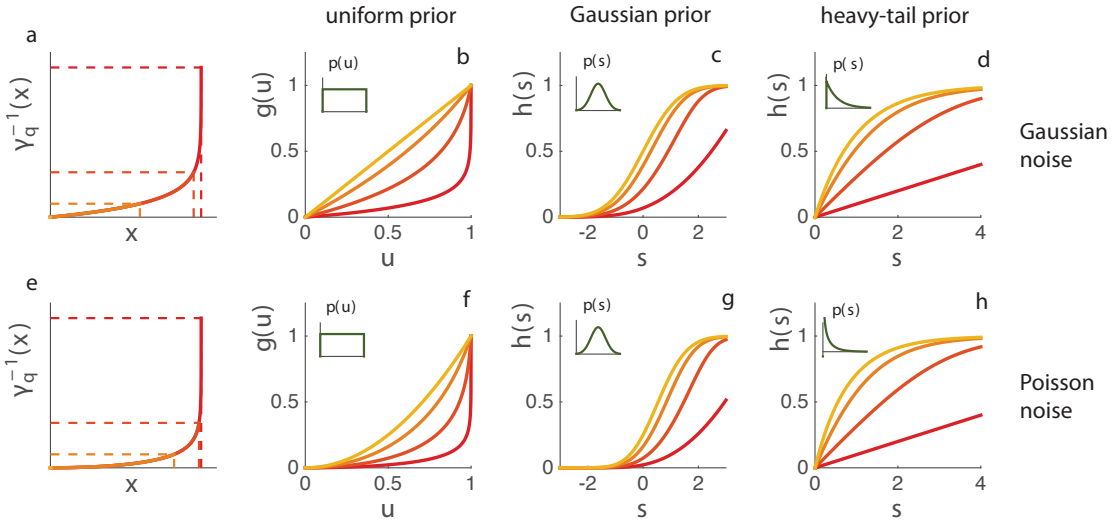

Figure 3: The process of determining optimal tuning curves $g(u)$ and corresponding $h(s)$ for different prior distributions and different noise models (top row: constant Gaussian noise $\alpha = 0$; bottom row: Poisson noise $\alpha = 1$). (a) A segment of the inverse incomplete gamma function is taken depending on the constraints. The higher the horizontal dash lines, the more substantial the metabolic constraint is. (b) The optimal $g(u)$ is determined for uniformly distributed u. (c) The corresponding optimal $h(s)$ for Gaussian prior. (d) The corresponding optimal $h(s)$ for Gamma distribution $p(s) \propto s^{q-1} \exp(-s)$. Specifically, in the absence of maximum response constraint and assuming the input follows this heavy tail distribution, the optimal tuning curve is exactly linear. (e-h) Similar to (a-d), but for Poisson noise.

most relevant situations.

Additive gaussian noise We begin with the simple additive Gaussian noise model, i.e. $\alpha = 0$. This model could provide a good characterization of the response mapping from the input stimulus to the membrane potential of a neuron (Laughlin, 1981). With more than sufficient metabolic supply, the optimal solution falls back to the principle of histogram equalization (see Figure 3b, yellow straight line). With less and less available metabolic budget, the optimal tuning curve bends downwards to satisfy this constraint. In general, the optimal solution strikes a balance between these two constraints, resulting in a family of optimal response functions in between of the two extrema mentioned above.

Poisson noise For neural spiking activity, it is observed that the variability often varies systematically with the mean firing rate (Tomko and Crapper, 1974; Tolhurst et al., 1981). In the case of Poisson spiking, the theory predicts the optimal response function should bend more downwards compared to the case of Gaussian noise (see Figure 3). In the extreme case when the main resource constraint comes from the limit firing rate range, the model predicts a square tuning curve $g(u) \propto u^2$ for uniform input (Figure 3f, yellow curve), which is consistent with early studies (Bethge et al., 2002; Johnson and Ray, 2004).

3.3.4. Distribution of the response magnitude

We also investigate the distribution of the response magnitude for the special case of linear metabolic costs ($\beta = 1$) and the result is summarized in Figure 4. In the case of Gaussian noise and K_{total} is large, the response magnitude is equally distributed in the response range. This is consistent with the histogram equalization solution which uses the response range equally well. However, as the metabolic constraint plays an increasingly important role when K_{total} is diminishing, the large response will be penalized more severely, resulting in more density at small response magnitude. We also found that Poisson noise leads to more penalization on large response magnitude compared to Gaussian noise, suggesting an interplay between noise and metabolic cost in shaping the optimal neural response distribution. Furthermore, in the case that K_{total} goes to 0, the response distribution converges to a gamma distribution, with heavy tail. This phenomenon represents the sparse codes (Olshausen and Field, 1996). It also gives a simple yet quantitative characterization of how the energy budget may push the neural responses toward a sparse coding regime.

3.4. Optimal Code for a Pair of Neurons

We next study the optimal coding in the case of two neurons with monotonic response functions. We denote the neural responses as $\mathbf{r} = (r_1, r_2)$. Therefore the efficient coding

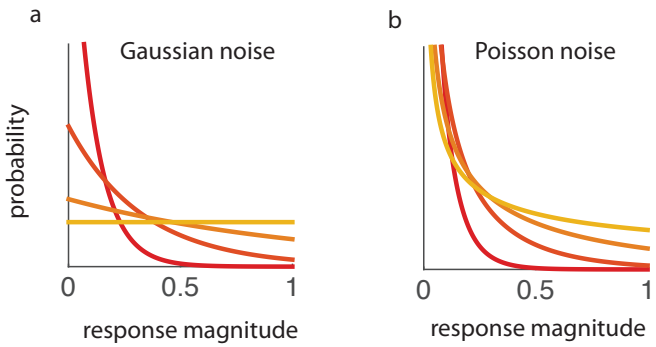

Figure 4: Distribution of the response based on the optimal response function of a single neuron. (a) Gaussian noise. (b) Poisson noise. In the extreme case of Gaussian noise with effectively no metabolic constraint, the distribution is uniformly distributed on the whole range.

problem becomes:

$$\begin{aligned} \text{maximize} \quad & L(\mathbf{h}) = \mathrm{MI}(s, \mathbf{r}) \\ \text{subject to} \quad & 0 \le h_i(s) \le r_{\max}, \quad i = 1, 2. && \text{(range constraint)} \\ & \mathbf{E}_s \left[K(h_1(s)) + K(h_2(s)) \right] \le 2K_{\text{total}} && \text{(metabolic constraint)} \end{aligned}$$

Note that we also double the available energy K_{total} so that on average, each neuron is still limited by same mean metabolic cost of K_{total} as it is for the single neuron case. Assuming the neural noise is independent across neurons, the system of two neurons has total Fisher information just as the linear sum of Fisher information contributed from each neuron $I_F(s) = I_1(s) + I_2(s)$.

3.4.1. Optimal response functions

Previous studies on neural coding with monotonic response functions have typically assumed that $h_i(s)$ has sigmoidal shape. It is important to emphasize that we do not make any a priori assumptions on the detailed shape of the tuning curve other than being monotonic and smooth. We define each neuron's active region $A_i = A_i^+ \cup A_i^-$ where $A_i^\pm = \{s | \pm h_i'(s) > 0\}$. Without going into detailed proof (see Section 3.8), we list the main conclusions

1. Neurons should have non-overlapping active regions $A_i \cap A_j = \emptyset$ if $i \neq j$.

2. If the metabolic constraint is binding, ON-OFF coding (A_1^+, A_2^- are non-empty or vice versa) is better than ON-ON coding (A_i^+'s are non-empty) or OFF-OFF coding (A_i^-'s are non-empty). Otherwise all three coding schemes can achieve the same mutual information.

3. For ON-OFF coding, it is better to have ON regions on the right side: $\sup A_i^- \leq \inf A_j^+$.

4. For ON-ON coding (or OFF-OFF), each neuron should have roughly the same tuning curve $h_i(s) \approx h_j(s)$ while still have disjoint active regions. Within the ON-pool or OFF-pool, the optimal tuning curve is same as the optimal solution from the single neuron case.

In Figure 5 we illustrate how these conclusions can be used to determine the optimal pair of neurons, assuming additive Gaussian noise $\alpha = 0$ and linear metabolic cost $\beta = 1$ (for other α, β, the process is similar). Crucially, our theory allows us to predict the precise shape of the optimal response functions. This represents a significant advantage over previous results on ON-OFF coding scheme using numerical methods (Karklin and Simoncelli, 2011) or restrictive neural codes (Gjorgjieva et al., 2014).

3.4.2. Comparison between ON-OFF and ON-ON codes

From Figure 5e we can see that, the maximum possible mutual information is monotonically increasing as a function of available energy K_{total} until they both saturate the limit at $K_{\text{ON-ON}} = 0.5 r_{\max}$ and $K_{\text{ON-OFF}} = 0.25 r_{\max}$ respectively (see the yellow tuning curves in Figure 5a-d). Note that these saturation limit is only valid for $\alpha = 0$ and $\beta = 1$. In order to encode exactly the same amount information, the most energy efficient ON-ON pair (or OFF-OFF) always requires twice as much compared to the most energy efficient ON-OFF pair.

On the other hand, we can compare the ON-ON and ON-OFF by fixing a value of $K_{\text{total}} < 0.5r_{\max}$ (i.e. when metabolic constraint is binding for ON-ON pairs). The optimal mutual information achieved by ON-ON neurons is always smaller than that achieved by ON-OFF neurons and the difference is plotted. If in the mutual information we use logarithm of base 2, this difference will saturate at -1 when the available energy is very limited $K_{\text{total}} \ll r_{\max}$. In this extreme case, the ON-ON code is only half as efficient as the ON-OFF code. In other words, it takes as much as twice amount of time T for the ON-ON code to achieve same amount of mutual information as the ON-OFF code.

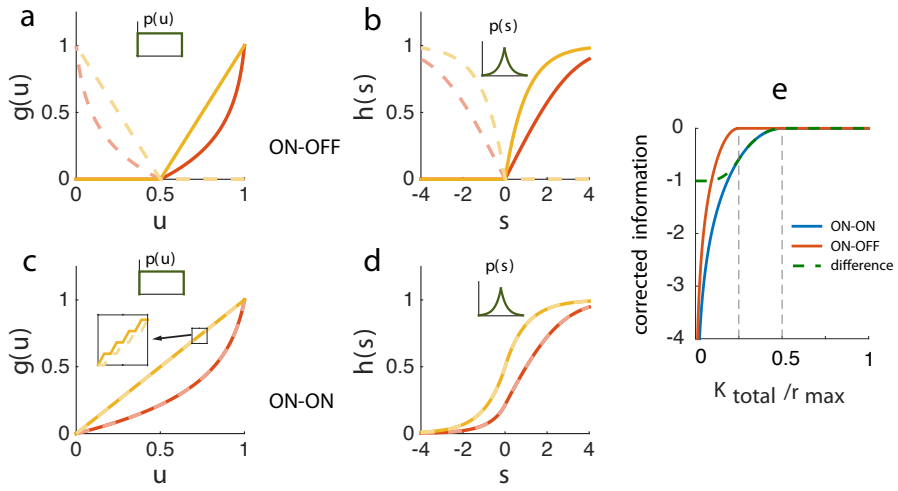

Figure 5: The optimal response functions for a pair of neurons assuming Gaussian noise. (a) The optimal response functions for a uniform input distribution assuming ON-OFF coding scheme. Solid yellow curve and dash yellow curve represent the optimal solution with weak metabolic constraint. Solid red and dash red curves are the optimal solution with substantial metabolic constraint. (b) Similar to panel a, but for input stimuli with heavy tail distribution. (c) The optimal response functions for a uniform input distribution assuming ON-ON coding scheme. Solid and dash yellow curves are for little metabolic constraint. Notice that two curves appear to be identical but are actually different at finer scales (see the inserted panel). Solid and dash red are for substantial metabolic constraint. (d) Similar to panel c, but for input stimuli with heavy tail distribution. (e) A comparison between the ON-ON scheme and ON-OFF scheme. The x-axis represents the relative importance of metabolic constraint. The y-axis represents the corrected information, defined as the amount of information actually transmitted minus the maximal information that can possibly be transmitted. The green dash line represent the difference between the information transmitted by the two schemes. Negative difference indicates an advantage of ON-OFF over ON-ON.

Our analysis provides a quantitative characterization of the advantage of ON-OFF over ON-ON and shows how it depends on the relative importance of the metabolic constraint. The encoding efficiency of ON-OFF ranges from double (when the metabolic budget is very limited) to equal amount of the ON-ON efficiency (when the metabolic budget exceeds certain threshold). This wide range includes the previous results where a mild increase (about 15%) is predicted in the efficiency when comparing ON-OFF to ON-ON under short integration time limit (Gjorgjieva et al., 2014). It is well known that in the retina of many animal species, there is a split of ON and OFF pathways (Schiller, 1992; Wässle, 2004). The substantial increase of efficiency in the regime of strong metabolic constraint supports the idea that strong metabolic constraint may be one of the main drives for such pathway splitting in evolution.

In a recent study by Karklin and Simoncelli (2011), it is observed numerically that training a simple linear-nonlinear network on natural images by maximizing mutual information subject to metabolic constraint would lead to ON-OFF coding scheme in certain noise regime. Our result may provide a theoretical bases for this observation, although we do not directly model the natural image, rather the neurons can been seen as encoding the local contrast in this context. Intriguingly, we found that in the case that the input distribution is a heavy tail distribution (see Figure 5b), the optimal response functions are two rectified non-linear functions who split the encoding range, which is similar to what has been observed physiologically in retina.

3.5. Optimal coding of large neural population

The framework could be generalized to study a large population of neurons ($N = 2k$, k is large). In this case, we consider the following problem:

$$
\begin{aligned}
\text{maximize} \quad & \mathrm{MI}(s, \mathbf{r}) \\
\text{subject to} \quad & r_{\min} \le h_i(s) \le r_{\max} && \text{(range constraint)} \\
& \mathbf{E}_s[\sum_i K(h_i(s))] \le N K_{\text{total}} && \text{(metabolic constraint)}
\end{aligned}
$$

We can again solve this problem analytically by exploiting the Fisher information approximation of mutual information (Brunel and Nadal, 1998; Wei and Stocker, 2016). Interestingly, we found the optimal codes should be divided into two pools of neurons of equal size k. One pool of neuron with monotonic increasing response function (ON-pool), and the other with monotonic decreasing response function (OFF-pool). For neurons within the same pool, the optimal response functions appear to be identical on the macro-scale but are quite different when zoomed in. We have shown that the optimal code must have disjoint active regions for each neuron. This is illustrated in the inset panel of Figure 5c, in which we show the case for two ON seemingly identical tuning curves.

We ask how the energy should be allocated across different neurons. Assume that metabolic cost is linear in terms of the response level and Poisson noise, each neuron across two different pools should share the same maximum firing rate. This generalizes to other noise type with considered ($\alpha > 0$) and other metabolic cost function ($\beta > 0$).

We quantify the amount of the information increase by using optimal coding schemes compared to using all ON neurons or all OFF neurons. Interestingly, the results we found in the Figure 5e for the a pair of neurons generalize to the current case. Specifically, in the case of strong metabolic constraint (*i.e.*, K_{total} is small), the optimal $2k$-ON neuron scheme is close to half of the efficiency of the optimal k-ON/k-OFF scheme.

The optimal coding scheme is reminiscent of the opponent coding observed in some neural systems, for example, the sound location system (Stecker et al., 2005). In our results the

support of the response function of an ON-neuron does not overlap with that of an OFF-neuron. We notice that in the physiological data (Stecker et al., 2005), there appears to be some overlap between two neuron which belong to different pools. However, in the case that there is noise in the input, it is possible that some amount of the overlap might be beneficial.

3.6. Discussion

We presented a theoretical framework for studying optimal neural codes under biologically relevant constraints. We emphasized the importance of two constraints – the noise characteristics of the neural responses and the metabolic cost. Throughout the paper, we have focused on neural codes with smooth monotonic response functions. We demonstrated that, maybe surprisingly, analytical solutions exist for a wide family of noise characteristics and metabolic cost functions.

An interesting venue for future research is to see whether the framework and techniques developed here could be used to define the optimal neural codes based on bell-shape tuning curves. Another interesting question is the optimal code in case of an odd number of neurons. Presumably, the solution for the case of $N = 2k + 1$ is close to $N = 2k$ for a large pool of neurons. However, when k is small, the difference due to symmetry breaking may substantially change the result. We have not addressed these results due to the lack of biological relevance for this case. Also, we have only considered the case of maximizing mutual information as the objective function; it will be interesting to see whether the results generalized to other objective functions such as, *e.g.*, minimizing decoding error (Twer and MacLeod, 2001; Wang et al., 2012).

Due to the limited scope of the paper, we have ignored several important other factors when formulating the efficient coding problem. First, we have not modeled the spontaneous activity of neurons, *i.e.* their baseline firing rate. Second, we have not considered the noise correlations between the responses of neurons. Third, we have ignored the noise in the

input to the neurons. We speculate that the first two factors are unlikely to significantly change our main results. However, incorporating the input noise may significantly change the results. In particular, for the cases of multiple neurons, our current results suggest that the response functions for ON and OFF neurons should not overlap. However, it is possible that this prediction does not hold in the presence of the input noise. Intuitively, introducing some redundancy by making the response functions partially overlap might be beneficial in this case. Including these factors into the framework should allow us to make a detailed and quantitative comparison to physiologically measured data in the future.

3.7. Appendix I: Determining constants a, b

In the main text we have showed that the optimal form of $g_*(u)$ is

$$g_*(u) = \left[\frac{1}{a}\gamma_q^{-1}\left(u\gamma_q(b)\right)\right]^{\frac{1}{\beta}} \tag{3.21}$$

where $q = (1-\alpha/2)/\beta$. The key part inside the bracket is a linearly scaled version of the inverse-incomplete-gamma function with two parameters a, b. Now we only need to re-write the constraints and the objective function in terms of a, b and find the optimal a, b that satisfies both constraints

3.7.1. Objective function

First we evaluate the objective function for each a, b

$$F(a,b) = \int_0^1 \log\left[\frac{g'(u)}{g(u)^{\alpha/2}}\right] du \tag{3.22}$$

where

$$g'(u) = \frac{1}{\beta}\left[\frac{1}{a}\gamma_q^{-1}(u\gamma_q(b))\right]^{\frac{1}{\beta}-1}\frac{1}{a}\left[\gamma_q^{-1}(u\gamma_q(b))\right]^{1-q}\exp(\gamma_q^{-1}(u\gamma_q(b)))\cdot\gamma_q(b)$$

$$= \frac{\gamma_q(b)}{\beta}\left(\frac{1}{a}\right)^{\frac{1}{\beta}}\left[\gamma_q^{-1}(u\gamma_q(b))\right]^{\frac{1}{\beta}-q}\exp(\gamma_q^{-1}(u\gamma_q(b))) \tag{3.23}$$

$$g(u)^{\alpha/2} = \left(\frac{1}{a}\right)^{\frac{\alpha}{2\beta}}\left[\gamma_q^{-1}(u\gamma_q(b))\right]^{\frac{\alpha}{2\beta}} \tag{3.24}$$

Since $q = (1-\alpha/2)/\beta$, we have

$$\frac{g'(u)}{g(u)^{\alpha/2}} = \frac{\gamma_q(b)}{\beta}a^{-q}\exp(\gamma_q^{-1}(u\gamma_q(b))) \tag{3.25}$$

$$F(a,b) = -\log\beta + \log\gamma_q(b) - q\log a + \int_0^1 \gamma_q^{-1}(u\gamma_q(b))\,du \tag{3.26}$$

Here we calculate the integral term. We let

$$v(u) = \gamma_q^{-1}(u\gamma_q(b)), \quad v(0) = 0, \quad v(1) = b. \tag{3.27}$$

$$u\gamma_q(b) = \gamma_q(v), \quad \gamma_q(b)\,du = v^{q-1}\exp(-v)\,dv \tag{3.28}$$

Therefore

$$\int_0^1 \gamma_q^{-1}(u\gamma_q(b))\,du = \frac{1}{\gamma_q(b)}\int_{v(0)}^{v(1)} v^q\exp(-v)\,dv \tag{3.29}$$

$$= \frac{1}{\gamma_q(b)}\left[q\int_0^b v^{q-1}\exp(-v)\,dv - v^q\exp(-v)|_0^b\right] \tag{3.30}$$

$$= \frac{1}{\gamma_q(b)}\left[q\gamma_q(b) - b^q\exp(-b)\right] = q - \frac{b^q\exp(-b)}{\gamma_q(b)} \tag{3.31}$$

Thus the objective function in terms of a, b is

$$F(a,b) = q - \log\beta + \log\gamma_q(b) - q\log a - \frac{b^q\exp(-b)}{\gamma_q(b)} \tag{3.32}$$

3.7.2. Optimal a for fixed value of b

We begin with rewriting the saturation constraint and the metabolic constraint in terms of a, b. First the saturation constraint

$$r_{\max} \geq g_*(1) = \left[\frac{b}{a}\right]^{\frac{1}{\beta}} \quad \Rightarrow \quad a \geq r_{\max}^{-\beta} \cdot b \stackrel{\text{def}}{=} A_1(b) \tag{3.33}$$

Second the metabolic constraint

$$\begin{aligned} K_{\text{ave}} \geq \int_0^1 K(g(u))\, du = \frac{1}{a}\int_0^1 \gamma_q^{-1}(u\gamma_q(b))\, du = \frac{1}{a}\left[q - \frac{b^q \exp(-b)}{\gamma_q(b)}\right] \\ \Rightarrow \quad a \geq K_{\text{ave}}^{-1} \cdot \left[q - \frac{b^q \exp(-b)}{\gamma_q(b)}\right] \stackrel{\text{def}}{=} A_2(b) \end{aligned} \tag{3.34}$$

Based on the form of the objective function $F(a, b)$, it is clear that a should be as small as possible, given that the above two constraints are satisfied. Therefore for fixed value of b, the smallest a that satisfies both constraints is

$$a_*(b) = \max\{A_1(b), A_2(b)\} \tag{3.35}$$

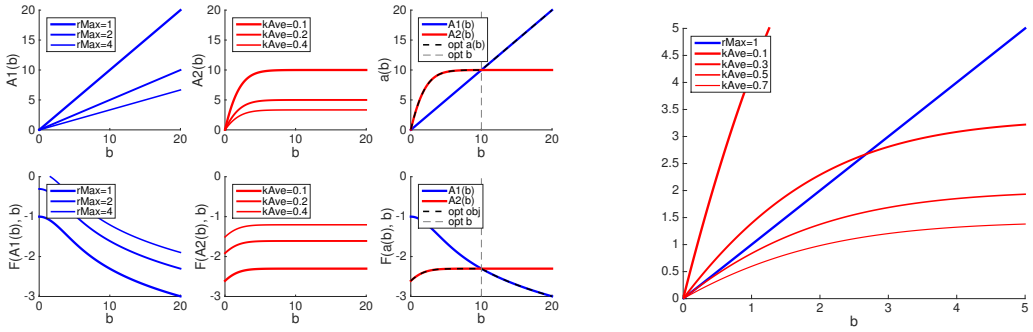

Figure 6: (LEFT) the plot of $A_1(b), A_2(b)$ and $a_*(b) = \max\{A_1(b), A_2(b)\}$ and corresponding objective function value. As we have proven, $F(A_1(b))$ is non-increasing and $F(A_2(b))$ is non-decreasing. The optimal value is achieved when $A_1(b) = A_2(b)$ if such solution b exists. (RIGHT) for fixed $r_{\max}$, the metabolic constraint can be redundant if the constant K_{ave} too large. When $\alpha = 0$, $\beta = 1$ (Gaussian noise, linear metabolic cost), the objective function can no longer be improved once K_{total} exceeds 0.5.

3.7.3. Optimal b

Here we discuss in cases. We consider two sets

$$B_1 = \{b \geq 0 | A_1(b) \geq A_2(b)\} \tag{3.36}$$

$$B_2 = \{b \geq 0 | A_2(b) \geq A_1(b)\} \tag{3.37}$$

Case I: For $b \in B_1$, $a \geq A_1(b)$ is the tighter constraint therefore $a_*(b) = A_1(b)$. Now we have

$$a_* = A_1(b) = r_{\max}^{-\beta} \cdot b \tag{3.38}$$

$$F(b) = F(a_*, b) = const + \log \gamma_q(b) - q \log b - \frac{b^q \exp(-b)}{\gamma_q(b)} \tag{3.39}$$

We will show that $F(a_*, b)$ is non-increasing in b. To prove this, we define an auxiliary function

$$Z(b) \stackrel{\text{def}}{=} \frac{b^q \exp(-b)}{\gamma_q(b)}, \quad \log Z(b) = q \log b - b - \log \gamma_q(b) \tag{3.40}$$

$$F(b) = const - \log Z(b) - b - Z(b) \tag{3.41}$$

We need to show

$$0 \geq F'(b) = -\frac{Z'(b)}{Z(b)} - Z'(b) - 1 \tag{3.42}$$

Here we calculate $Z'(b)$

$$Z'(b) = \frac{q b^{q-1} \exp(-b) - b^q \exp(-b)}{\gamma_q(b)} - \frac{b^{2q-1} \exp(-2b)}{\gamma_q(b)^2} \tag{3.43}$$

$$= \frac{b^{q-1} \exp(-b)}{\gamma_q(b)^2} \underbrace{[(q-b)\,\gamma_q(b) - b^q \exp(-b)]}_{\stackrel{\text{def}}{=} Z_2(b)} \tag{3.44}$$

The term $Z_2(b)$ has property

$$Z_2(0) = 0, \quad Z_2'(b) = -\gamma_q(b) \quad \Rightarrow \quad Z_2(b) = -\int_0^b \gamma_q(t)\, dt \tag{3.45}$$

Therefore

$$Z'(b) = -\underbrace{\frac{b^q \exp(-b)}{\gamma_q(b)}}_{=Z(b)} \underbrace{\frac{\int_0^b \gamma_q(t)\, dt}{b\gamma_q(b)}}_{=M(b)} = -Z(b)M(b) \tag{3.46}$$

Plug this into $F'(b)$

$$F'(b) = M(b)(1 + Z(b)) - 1 \tag{3.47}$$

First we can show $F'(0) = 0$. Now for $b > 0$, we have

$$F_2(b) \stackrel{\text{def}}{=} F'(b) \cdot b\gamma_q(b) = \int_0^b \gamma_q(t)\, dt \cdot (1 + Z(b)) - 1 \tag{3.48}$$

$$F_2'(b) = \gamma_q(b)\,(1 + Z(b)) + \int_0^b \gamma_q(t)\, dt \cdot Z'(b) - \gamma_q(b) - b^q \exp(-b) = \int_0^b \gamma_q(t)\, dt \cdot Z'(b) < 0 \tag{3.49}$$

Therefore $F'(b) \leq 0$ and the function $F(b)$ is non-increasing. This means that in the case of $b \in B_1$, the optimal solution is the smallest $b_* = \inf_{b \in B_1} b$.

Case II: For $b \in B_2$, $a \geq A_2(b)$ is the tighter constraint therefore $a_*(b) = A_2(b)$. Now we have

$$a_*(b) = A_2(b) = K_{\text{ave}}^{-1}\left[q - \frac{b^q \exp(-b)}{\gamma_q(b)}\right] = K_{\text{ave}}^{-1}\left[q - Z(b)\right] \tag{3.50}$$

$$F(b) = F(a_*, b) = const + \log \gamma_q(b) - q \log(q - Z(b)) - Z(b) \tag{3.51}$$

Now we will show this $F(b)$ is non-decreasing in b.

$$F'(b) = \frac{b^{q-1}\exp(-b)}{\gamma_q(b)} + \frac{qZ'(b)}{q - Z(b)} - Z'(b) \tag{3.52}$$

$$= \frac{Z(b)}{b} + Z'(b)\frac{Z(b)}{q - Z(b)} = \frac{Z(b)}{b(q - Z(b))}\left[q - Z(b) + bZ'(b)\right] \tag{3.53}$$

The term outside the bracket is positive when $b > 0$. Therefore we only need to show

$$Z_3(b) = q - Z(b) + bZ'(b) \geq 0. \tag{3.54}$$

Note that we have

$$\log Z(b) = q\log b - b - \log\gamma_q(b) \tag{3.55}$$

$$\frac{Z'(b)}{Z(b)} = \frac{q}{b} - 1 - \frac{b^{q-1}\exp(-b)}{\gamma_q(b)} = \frac{q}{b} - 1 - \frac{Z(b)}{b} \tag{3.56}$$

Therefore

$$q - Z(b) = b\left(1 + \frac{Z'(b)}{Z(b)}\right) \tag{3.57}$$

Plug this into Eq(57) we have

$$Z_3(b) = b\left(1 + \frac{Z'(b)}{Z(b)} + Z'(b)\right) \geq 0 \tag{3.58}$$

which share the proof of the inequality in [Eq(45)]. In this case, $F(b)$ is non-decreasing. This means that in the case of $b \in B_2$, the optimal solution is the largest $b_* = \sup_{b\in B_2} b$.

Conclusion: Based on these two cases, we know that if both B_1, B_2 are non-empty, then the optimal b_* is both the infimum or B_1 and supremum of B_2, which means that b_* is

uniquely determined by

$$A_1(b) = \frac{1}{r_{\max}^{\beta}} b = \frac{1}{K_{\text{ave}}} \left[q - \frac{b^q \exp(-b)}{\gamma_q(b)} \right] = A_2(b) \tag{3.59}$$

Since $A_1(b)$ grows linearly but $A_2(b)$ has an upper bound, therefore B_1 cannot be empty. However if B_2 is empty, then the optimal $b_* = \inf_{b \in \mathbf{R}^+} b = 0$ which means that the optimal solution is attained by the limit $b \to 0$.

3.8. Appendix II: Technical Details for Multiple Neurons Case

First we define the active regions of the i-th neuron as

$$A_i^{\pm} = \{s | \pm h_i'(s) > 0\}, \quad A_i = A_i^- \cup A_i^+. \tag{3.60}$$

Now we prove a couple of necessary conditions for these A_i to be optimal in terms of maximum mutual information. Note that the tuning curves are assumed to be monotonic so one of A_i^+ and A_i^- must be empty.

As a useful preliminary result, we recall that the total Fisher information of the population is the linear sum of Fisher information contributed by each individual neuron

$$I_F(s) = \sum_{i=1}^{N} I_i(s) \quad \text{where} \quad I_i(s) \propto \frac{h_i'(s)^2}{h_i(s)^{\alpha}} \tag{3.61}$$

if the noise model parameter is α. It is clear that $I_i(s)$ is greater than zero only if $s \in A_i$.

Lemma 3.8.1 (Non-overlapping active regions.)**.** *We consider the problem of optimizing a neural population with neuron $i = 1, \ldots, N$. We limit the stimulus to be on some subset $s \in [s_0, s_1]$ of the original range $[0, 1]$. Each neuron is monotonic (either $h_i'(s) \geq 0$ or*

$h_i'(s) \leq 0$ *for* $s \in [s_0, s_1]$*) and has limited range of output* $L_i \leq h_i(s) \leq H_i$*.*

$$\textit{maximize} \quad \int_{s_0}^{s_1} \log I_F(s)\, ds \tag{3.62}$$

$$\textit{subject to} \quad L_i \leq h_i(s) \leq H_i, \quad i = 1, \ldots, N \tag{3.63}$$

Then a necessary condition is the ***non-overlapping active regions****, i.e.* $A_i \cap A_j = \emptyset$ *for all* $i \neq j$*.*

Proof. We begin with the proof of an upper bound on the integral of the square root of the Fisher information:

$$\int_{s_0}^{s_1} \sqrt{I_i(s)}\, ds \leq I_i^{\max} \tag{3.64}$$

For $\alpha \neq 2$, for example, we have

$$\sqrt{I_i(s)} \propto \frac{|h_i'(s)|}{h_i(s)^{\alpha/2}} \propto \frac{d}{ds}\left[h_i(s)^{1-\alpha/2}\right] \tag{3.65}$$

$$I_i^{\max} \propto |H_i^{1-\alpha/2} - L_i^{1-\alpha/2}| \tag{3.66}$$

For $\alpha = 2$ the calculation is similar if we use $\log h_i(s)$ as the anti-derivative. Next we write down an upper bound of the objective function:

$$I_F(s) = \sum_{i=1}^{N} I_i(s) = \left(\sum_{i=1}^{N} \sqrt{I_i(s)}\right)^2 - 2\sum_{i<j} \sqrt{I_i(s) \cdot I_j(s)} \leq \left(\sum_{i=1}^{N} \sqrt{I_i(s)}\right)^2 \stackrel{\text{def}}{=} Q(s)^2 \tag{3.67}$$

$$\int_{s_0}^{s_1} \log I_F(s)\, ds \leq 2\int_{s_0}^{s_1} \log Q(s)\, ds = 2(s_1 - s_0)\int_{s_0}^{s_1} \frac{1}{s_1 - s_0} \log Q(s)\, ds \tag{3.68}$$

$$\leq 2(s_1 - s_0) \log \int_{s_0}^{s_1} \frac{Q(s)}{s_1 - s_0}\, ds \leq 2(s_1 - s_0) \log \frac{\sum_i I_i^{\max}}{s_1 - s_0}. \tag{3.69}$$

where we have used the Jensen's inequality and the optimization constraints. To achieve this attainable upper bound for the objective function, we need $Q(s) = const$ for the Jensen's

inequality and also the equality in Eq. 3.67. Therefore a necessary condition for $h_i(s)$ to be optimal is that $I_i(s) \cdot I_j(s) = 0$ everywhere for $i \neq j$. This is equivalent as our claim $A_i \cap A_j = \emptyset$ for all $i \neq j$. □

In other words, different neurons should not have non-overlapping active region. However, the above lemma does not take the energy constraints into consideration. If we add the energy constraint

$$\int_0^1 \sum_{i=1}^{N} K(h_i(s))\, ds \leq K_{\text{total}} \tag{3.70}$$

does it break the necessity of the non-overlapping Fisher information condition? The answer is no due to the following lemma.

Lemma 3.8.2 (Non-overlapping active regions with metabolic constraints)**.** *Assuming $h_i(s)$ is the optimal solution to the following problem. Each neuron is monotonic (either $h_i(s) \geq 0$ or $h_i(s) \leq 0$ for $s \in [0,1]$) and has limited range of output $L \leq h_i(s) \leq H$.*

$$\textit{maximize} \quad \int_0^1 \log I_F(s)\, ds \tag{3.71}$$

$$\textit{subject to} \quad L \leq h_i(s) \leq H, \quad i = 1, \ldots, N \tag{3.72}$$

$$\int_0^1 \sum_{i=1}^{N} K(h_i(s))\, ds \leq K_{total} \tag{3.73}$$

Then a necessary condition is the ***non-overlapping active regions****, i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$.*

Proof. We show this lemma by contradiction – we assume $h_i(s)$ is optimal with $I_i(s) \cdot I_j(s) > 0$ for some s (so $s \in A_i \cap A_j$) and show that there exists a better solution $\tilde{h}_i(s)$.

We divide the stimulus space $s \in [0,1]$ equally into M smaller intervals with endpoints

$s_j = j/M$ for $j = 0, \ldots M$. We define

$$L_{i,j} = \min_{s \in [s_{j-1}, s_j]} h_i(s) \tag{3.74}$$

$$H_{i,j} = \max_{s \in [s_{j-1}, s_j]} h_i(s) \tag{3.75}$$

On each of these intervals $[s_{j-1}, s_j]$ and the above range, we apply Lemma 1 and obtain a new solution $\tilde{h}_i(s)$ which satisfies the non-overlapping Fisher information condition. It is easy to see that this new solution gives better objective function. Next we show that this better solution costs similar amount of energy as $h_i(s)$. Using the upper and lower bound of firing rate in each interval, we have

$$\int_0^1 K(\tilde{h}_i(s))\, ds = \sum_{j=1}^{M} \int_{s_{j-1}}^{s_j} K(\tilde{h}_i(s))\, ds \leq \sum_{j=1}^{M} (s_j - s_{j-1}) K(H_{i,j}) = \frac{1}{M} \sum_{j=1}^{M} K(H_{i,j}) \tag{3.76}$$

$$\int_0^1 K(\tilde{h}_i(s))\, ds = \sum_{j=1}^{M} \int_{s_{j-1}}^{s_j} K(\tilde{h}_i(s))\, ds \geq \sum_{j=1}^{M} (s_j - s_{j-1}) K(L_{i,j}) = \frac{1}{M} \sum_{j=1}^{M} K(L_{i,j}) \tag{3.77}$$

Similarly for the original solution $h_i(s)$ these two bounds also apply

$$\frac{1}{M} \sum_{j=1}^{M} K(L_{i,j}) \leq \int_0^1 K(h_i(s))\, ds \leq \frac{1}{M} \sum_{j=1}^{M} K(H_{i,j}) \tag{3.78}$$

Therefore

$$|K_{\tilde{h}} - K_h| = \left| \int_0^1 K(\tilde{h}_i(s))\, ds - \int_0^1 K(h_i(s))\, ds \right| \tag{3.79}$$

$$\leq \frac{1}{M} \sum_{j=1}^{M} (H_{i,j} - L_{i,j}) = \frac{1}{M}(H - L) \tag{3.80}$$

and the right side converges to zero as M goes to infinity. This means that the energy consumption of this new solution $\tilde{h}_i(s)$ can be made as close to the original solution as possible if we use a finer and finer grid (large M), while having a better objective function value. This contradicts the optimality of $h_i(s)$. □

Using Lemma 1 and Lemma 2, we conclude that in the optimal population, the neurons need to have non-overlapping Fisher information. This simplifies our further analysis. Here we discuss another necessary condition that a pair of ON-OFF neurons must satisfies.

Lemma 3.8.3 (ON-OFF neurons)**.** *Under the assumption that the energy constraint is binding, then for an ON-neuron with active region A_i^+ and an OFF-neuron with active region A_j^-, we have $\sup B_j^- \leq \inf A_i^+$. In other words, the active region of any ON neuron is strictly on the right side of the active region of any OFF neuron.*

Proof. We denote $s_i = \inf A_i^+$ and $s_j = \sup A_j^-$ and prove the lemma by contradiction. We assume $s_i < s_j$. Due to the piecewise continuity of $h_i'(s)$, there exists $\epsilon > 0$ such that there exist small neighborhoods $[s_i, s_i + \epsilon] \in A_i^+$ and $[s_j - \epsilon, s_j] \in A_j^-$. We can construct a new tuning curve $\tilde{h}_i(s)$ and $\tilde{h}_j(s)$ by swapping their active regions (see Figure 7 below) in these neighborhood. It is obvious that the new $\tilde{h}_i$ costs strictly less amount of energy. The

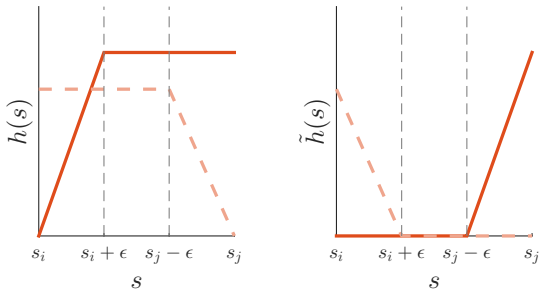

Figure 7: Tuning curve surgery for ON-OFF neuron pairs that reduces energy costs.

new tuning curves has equal performance in terms of objective function because the regions being swapped has same size. The existence of such tuning curves contradicts the fact that the energy constraint is binding. □

One immediate corollary is that, in a large population of neurons with both ON and OFF sub-populations, there exists a single s that divides the active regions for the ON sub-population and the OFF sub-population.

Next we find the optimal condition for a population of only ON/OFF neurons. Without

loss of generality, we assume the neural population consists of only ON neurons.

Lemma 3.8.4 (ON-ON neurons)**.** *Assuming the population has only ON neurons and the metabolic cost function is linear* $\sum_i K(h_i(s)) = K(\sum_i h_i(s))$*, then the optimal* $h_i(s) \approx h(s)/N$ *but with disjoint active regions* A_i^+*. The function* $h(s) = \sum_i h_i(s)$ *is the single neuron infomax solution of:*

$$\begin{aligned} \textit{maximize}_{h(s)} \quad & \mathrm{MI}(s, r) && (3.81)\\ \textit{subject to} \quad & 0 \le h(s) \le N \cdot r_{max} && (3.82)\\ & \mathbf{E}[K(h(s))] \le N \cdot K_{total} && (3.83) \end{aligned}$$

Proof. We denote $h_i(s) = p_i(s) \cdot h(s)$ and it is clear that $\sum p_i(s) = 1$. Using Lemma 1 we know A_i's are disjoint therefore we also have $h_i'(s) = h'(s) \cdot 1_{A_i}$. Plug these into the objective function, we know that

$$I_F(s) \propto \sum_{i=1}^{N} \frac{h_i'(s)^2}{h_i(s)^\alpha} = \frac{h'(s)^2}{h(s)^\alpha} \cdot \sum_{i=1}^{N} 1_{A_i} \cdot p_i(s)^{-\alpha} \quad \Rightarrow \tag{3.84}$$

$$\int_0^1 \log I_F(s)\, ds = \int_0^1 \log \frac{h'(s)^2}{h(s)^\alpha}\, ds + \int_0^1 \log \left(\sum_{i=1}^{N} 1_{A_i} \cdot p_i(s)^{-\alpha} \right) ds \tag{3.85}$$

Now the problem is divided into two independent part. The first part involves finding the optimal $h(s)$ under constraints stated in the lemma. This part is exactly the same as the single neuron case.

The second part involves optimizing A_i and $p_i(s)$ for the following term

$$\begin{aligned} \text{maximize} \quad & \int_0^1 \log \left(\sum_{i=1}^{N} 1_{A_i} \cdot p_i(s)^{-\alpha} \right) ds = -\alpha \sum_i \int_0^1 1_{A_i} \log p_i(s)\, ds && (3.86)\\ \text{subject to} \quad & \sum p_i(s) = 1. && (3.87) \end{aligned}$$

Using Lagrange multiplier method we know the optimal condition for $p_i(s)$ is

$$-\alpha \frac{1_{A_i}}{p_i(s)} - \lambda = 0 \tag{3.88}$$

This shows that $p_i(s) = const$ for $s \in A_i$. However A_i is the active region of neuron i so the function $h_i(s)$ is increasing and all other $h_j(s)$ remains the same. The only way for this condition to holds is when A_i consists of infinite many small intervals so that the increase in $h_i(s)$ is small. Also we know that all $p_i(s) \to 1/N$ when $s \to 1$. Therefore one possible solution is given by $h_i(s) \approx h(s)/N$ but on infinitesimal scales, each small interval is equally divided into N disjoint set A_i's. □

CHAPTER 4 : L_p Optimal Codes for One Dimensional Stimulus

4.1. Introduction

The efficient coding hypothesis states that biological sensory systems have limited coding resources and therefore seek to employ coding strategies that are optimally adapted to the statistical structure of their sensory environment (Attneave, 1954; Barlow, 1961; Maddess and Laughlin, 1985; Theunissen and Miller, 1991; Fitzpatrick et al., 1997; Harper and McAlpine, 2004). Several studies have experimentally demonstrated that sensory neural codes seem to indeed follow input distribution statistics in order to reach higher coding efficiency (Brenner et al., 2000; Twer and MacLeod, 2001; Dean et al., 2005; Ozuysal and Baccus, 2012). A large fraction of previous work assumed that neural representations are tuned to maximize the mutual information they are able to convey about the stimulus values given some overall constraints on available metabolic costs, *e.g.* total number of spikes (Laughlin, 1981; Linsker, 1989; Atick. and Redlich, 1990; Van Hateren, 1993; Seung and Sompolinsky, 1993; Nadal and Parga, 1994; Brunel and Nadal, 1998; Zhang and Sejnowski, 1999; Pouget et al., 1999; Kang et al., 2004; Sharpee et al., 2006; McDonnell and Stocks, 2008; Nikitin et al., 2009; Tkacik et al., 2010; Yarrow et al., 2012; Kastner et al., 2015). This *Infomax* criterion has been a preferred choice because it does not require making any further assumptions about potential downstream computations and tasks the encoded stimulus may be involved in. On the other hand, a few studies have taken a downstream perspective and have argued for optimality criteria that consider how well the stimulus information can actually be reconstructed from the neural representations. They often use a metric criterion in terms of the mean squared reconstruction error (Bethge et al., 2002, 2003; Berens et al., 2009; Yaeli and Meir, 2010; Doi and Lewicki, 2011). This reconstruction metric has been shown to optimize performance in perceptual estimation and classification tasks (Salinas, 2006). Recently there have been increasing interest in comparing the information with the metric approach (Ganguli and Simoncelli, 2010; Gjorgjieva et al., 2014; Grabska-Barwinska and Pillow, 2014). However, a unified comparison and evaluation of

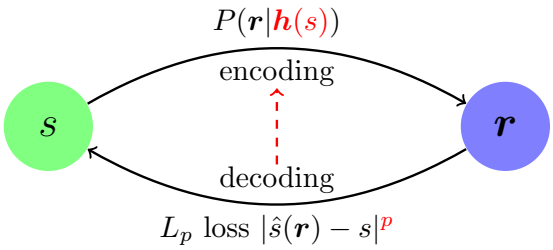

Figure 8: Efficient coding problem in terms of reconstruction error. A one-dimensional stimulus s is encoded in a neural response pattern r. We define the optimal tuning curve $\boldsymbol{h}(s)$ as the one that minimizes the overall L_p reconstruction error according to an MLE decoder. We study how the optimal coding strategy is dependent on the norm parameter p. The Infomax solution is equivalent to the optimal encoder for $p \to 0$.

these different approaches is currently lacking.

Here, we provide a unified framework to compare these optimal criteria. We introduce a parametric formulation of the efficient coding problem in terms of minimizing the overall reconstruction error according to the L_p norm, as a function of the norm parameter p. We assume reconstruction from a maximum likelihood estimate (MLE) decoder in the asymptotic time limit. More specifically, we consider a one-dimensional stimulus s with distribution $f(s)$ that is encoded with tuning curve(s) $\boldsymbol{h}(s)$ for m neuron(s). While the mapping $\boldsymbol{h}(s)$ is deterministic, we assume the neural response $\mathbf{r}$ to follow a distribution $P(\mathbf{r}|\boldsymbol{h}(s))$ according to neural noise. For both Poisson and Gaussian noise, we analytically derive the optimal tuning curve $\boldsymbol{h}$ to achieve minimal L_p mean reconstruction error for arbitrary stimulus distributions. This framework includes both the Infomax as well as mean-squared error optimal solutions in the limit of $p \to 0$ and $p = 2$ respectively. We first focus on solutions for the optimal tuning curve $h(s)$ of a single (sigmoidal) neuron encoding the stimulus. We then show how the single neuron tuning curve solution can be naturally extended to populations of neurons. Under certain assumptions, the optimal single neuron tuning curve $h(s)$ can be related to an optimal *meta-tuning curve* of the neural population, from which the individual tuning characteristics of the population of neurons can be determined.

In the context of this theoretical framework, we investigate how known tuning characteristics of biological sensory systems can be explained. We compare the measured tuning characteristics of early sensory representations in the fly, the cat, and the monkey for known stimulus statistics with predictions from our framework. For the examples we tested, the biological tuning characteristics are quite well predicted by our framework, and are best matched for small values of the norm parameter p. We conclude that early sensory representations in biological systems may be optimized to convey maximal information.

4.2. Optimal Neural Coding for a Single Neuron

We start with the case where a single neuron is encoding a one-dimensional stimulus variable s. We assume that s follows a distribution density $f(s)$. We also assume that the neuron's average firing rate is determined by a sigmoidal function $h(s)$. The actual observed firing rate r is subject to neural noise, whose variability is described by a stochastic model $P(r|h(s))$.

We do not limit the noise to be defined by canonical Poisson spiking model. Rather, we only assume that (a) the mean firing rate is equal to the output of the tuning curve $\langle r \rangle = h(s)$ and (b) the spike generating process is independent from the neuron's spiking history. With sufficient encoding time or with independent observations of identical neurons, the accumulated noise is asymptotically normal with zero mean and fixed variance according to the Central Limit Theorem. In order to decode the input stimulus s, we take the maximum likelihood estimator (MLE) $\hat{s}(r)$, which is asymptotically unbiased and efficient (Cover and Thomas, 1991).

In order to find the L_p optimal tuning curve for a one dimensional stimulus s, we need to minimize the mean L_p loss of the maximum likelihood estimator. The only constraint for the sigmoidal tuning curve is the saturation limits of the firing rates. Within the regime of low noise limit, the maximum firing rate does not affect the optimality. Therefore we

assume $0 \leq h(s) \leq 1$ without loss of generality, which leads to the optimization problem

$$\text{minimize} \qquad \left\langle |\hat{s}(r) - s|^p \right\rangle_{s,r} \tag{4.1}$$

$$\text{subject to} \qquad 0 \leq h(s) \leq 1. \tag{4.2}$$

4.2.1. Objective Functions in terms of Fisher Information

In this chapter we need to use the concept of Fisher information intensively. The Fisher information $\mathcal{I}(s)$ describes the precision of the best possible estimator for each specific individual stimulus s. In case of a one dimensional input s, $\mathcal{I}(s)$ can be calculated according to its definition

$$\mathcal{I}(s) = \left\langle \left(\frac{\partial}{\partial s} \log p(r|s) \right)^2 \middle| s \right\rangle_{p(r|s)} \tag{4.3}$$

where the conditional distribution $p(r|s)$ describes the stochastic neural response for a given stimulus and the average is taken over r but not s. It has been shown that in the asymptotic limit of long encoding time, the total Fisher information characterizes the precision of the estimator $\hat{s}$ in reconstructing the stimulus s (see Chapter 2)

$$(\hat{s}(r) - s) \sim \text{Normal}(0, \mathcal{I}(s)^{-1}) \tag{4.4}$$

$$\left\langle |\hat{s}(r) - s|^p \middle| s \right\rangle_r = const(p) \cdot \mathcal{I}(s)^{-p/2} \tag{4.5}$$

It is clear from Eq. (4.5) that larger Fisher information leads to smaller L_p error. One example is the Cramer-Rao lower bound when $p = 2$. The more general Eq. (4.5) establishes the connection between L_p loss in Eq. (4.1) and the Fisher information. This leads to an equivalent optimization in terms of Fisher information:

$$\text{minimize} \qquad \left\langle \mathcal{I}(s)^{-p/2} \right\rangle_s \tag{4.6}$$

In addition to the L_p-error minimization problem, we also consider the well-known Infomax optimization which maximizes the mutual information between the response and the stimulus. It has previously been shown that Fisher information can be related to mutual information (Brunel and Nadal, 1998; Wei and Stocker, 2016). In our framework, Infomax is equivalent to optimizing the logarithm of Fisher information:

$$\mathrm{MI}(\mathbf{r}, s) = \frac{1}{2} \left\langle \log \mathcal{I}(s) \right\rangle_s + const \tag{4.7}$$

$$\text{minimize} \qquad - \left\langle \log \sqrt{\mathcal{I}(s)} \right\rangle_s \tag{4.8}$$

4.2.2. Constraints in terms of Fisher Information

Next we show how to incorporate constraints in Eq. (4.2) into the same framework. For a one dimensional stimulus variable, the Fisher information of a neuron is fully determined by the nonlinear tuning curve $h(s)$ and the noise model. Here we show the results for Poisson noise (P), constant Gaussian noise (cG) and generalized Gaussian noise (gG).

$$\textbf{P:} \quad \mathcal{I}(s) \propto T \cdot \frac{h'(s)^2}{h(s)} \tag{4.9}$$

$$\textbf{cG:} \quad \mathcal{I}(s) \propto T \cdot h'(s)^2 + O(1) \tag{4.10}$$

$$\textbf{gG:} \quad \mathcal{I}(s) \propto T \cdot \frac{h'(s)^2}{h(s)^\alpha} + O(1) \tag{4.11}$$

In the asymptotic long time limit $T \to \infty$, these formulae can easily be inverted – for any given Fisher information allocation $\mathcal{I}(s)$, the corresponding nonlinear tuning curve is

$$\textbf{P:} \quad h(s) \propto \left(\int_{-\infty}^{s} \sqrt{\mathcal{I}(\xi)}\, d\xi \right)^2 \tag{4.12}$$

$$\textbf{cG:} \quad h(s) \propto \int_{-\infty}^{s} \sqrt{\mathcal{I}(\xi)}\, d\xi \tag{4.13}$$

$$\textbf{gG:} \quad h(s) \propto \left(\int_{-\infty}^{s} \sqrt{\mathcal{I}(\xi)}\, d\xi \right)^{1/(1-\alpha/2)} \tag{4.14}$$

Given bound constraints on the tuning curve in Eq. (4.2), we have

$$\textbf{P:} \quad \int_{-\infty}^{\infty} \sqrt{\mathcal{I}(s)}\, ds \propto \int_{-\infty}^{\infty} \frac{h'(s)}{\sqrt{h(s)}}\, ds = 2\sqrt{h(s)}\Big|_{-\infty}^{\infty} \leq const \tag{4.15}$$

$$\textbf{cG:} \quad \int_{-\infty}^{\infty} \sqrt{\mathcal{I}(s)}\, ds \propto \int_{-\infty}^{\infty} h'(s)\, ds = h(s)|_{-\infty}^{\infty} \leq const \tag{4.16}$$

$$\textbf{gG:} \quad \int_{-\infty}^{\infty} \sqrt{\mathcal{I}(s)}\, ds \propto \int_{-\infty}^{\infty} \frac{h'(s)}{h(s)^{\alpha/2}}\, ds = \frac{1}{1-\alpha/2} h(s)^{1-\alpha/2}\Big|_{-\infty}^{\infty} \leq const \tag{4.17}$$

Ignoring irrelevant constant scalar terms which do not affect the optimal form, these constraints can be unified as a single constraint on the integral of the square root of Fisher information:

$$\textbf{P, cG or gG:} \quad \text{subject to} \qquad \int_{-\infty}^{\infty} \sqrt{\mathcal{I}(s)}\, ds \leq const \tag{4.18}$$

Since it is always better to have more Fisher information, equality in Eq. (4.18) must hold for optimality. To summarize, the objective function in Eq. (4.6) attempts to optimally allocate the Fisher information $\mathcal{I}(s)$ across the space of the stimulus variable s with distribution $f(s)$ under the integral constraint in Eq. (4.18). After determining the optimal allocation $\mathcal{I}^*(s)$, the optimal nonlinearity $h^*(s)$ can then be determined using Eq. (4.12), Eq. (4.13) or Eq. (4.14), depending upon the neural noise model.

4.2.3. Single Neuron Results

According to the above analysis, solving the L_p reconstruction error minimization problem is equivalent to solving the Fisher information allocation problem. For each p value in the L_p-minimum decoding loss criterion, the optimization problem is

$$\text{minimize} \qquad \left\langle (\mathcal{I}(s))^{-p/2} \right\rangle_s = \int f(s)\, (\mathcal{I}(s))^{-p/2}\, ds \tag{4.19}$$

$$\text{subject to} \qquad \int \sqrt{\mathcal{I}(s)}\, ds \leq const \tag{4.20}$$

This variational problem can easily be solved and the optimal solution is

$$\mathcal{I}^*(s) \propto f(s)^{2/(1+p)} \tag{4.21}$$

$$\textbf{P:} \qquad h^*(s) = \left(\frac{\int_{-\infty}^{s} f(\xi)^{1/(1+p)}\, d\xi}{\int_{-\infty}^{\infty} f(\xi)^{1/(1+p)}\, d\xi} \right)^2 \tag{4.22}$$

$$\textbf{cG:} \qquad h^*(s) = \frac{\int_{-\infty}^{s} f(\xi)^{1/(1+p)}\, d\xi}{\int_{-\infty}^{\infty} f(\xi)^{1/(1+p)}\, d\xi} \tag{4.23}$$

$$\textbf{gG:} \qquad h^*(s) = \left(\frac{\int_{-\infty}^{s} f(\xi)^{1/(1+p)}\, d\xi}{\int_{-\infty}^{\infty} f(\xi)^{1/(1+p)}\, d\xi} \right)^{1/(1-\alpha/2)} \tag{4.24}$$

A simple comparison between the two noise models reveals that the optimal tuning curve for a neuron with Poisson noise is exactly the square of the optimal tuning curve for a neuron with constant Gaussian noise. This relationship was first reported by (Bethge et al., 2002) and (Johnson and Ray, 2004). The squaring transformation shows that the optimal coding under Poisson noise tends to utilize more reliable low firing rates rather than more unreliable higher rates. Below we focus on the constant Gaussian noise solution and discuss the link between our general formula and several results that have been previously reported in the literature:

- When $p = 0$, the L_0-minimum solution is given by the cumulative function of the input distribution,

$$h^*(s) \propto \int_{-\infty}^{s} f(\xi)\, d\xi. \tag{4.25}$$

- When $p = 2$, the L_2-minimum solution is given by the cumulative function of the cube root of the input distribution,

$$h^*(s) \propto \int_{-\infty}^{s} f(\xi)^{1/3}\, d\xi \tag{4.26}$$

- When $p \to \infty$, the optimal tuning curve $h^*(s)$ converges to a linear function because

its derivative approaches a constant function of s and the prior $p(s)$ is no longer relevant. However this usually requires the stimulus to be bounded $s \in [s_{\min}, s_{\max}]$ otherwise the integral of $f(s)^{1/(1+p)}$ will diverge for sufficiently large p.

Note that optimizing the L_p-min problem Eq. (4.19) when $p \to 0$ leads to the same optimal solution as the infomax problem in Eq. (4.8). This solution, first proposed in (Laughlin, 1981; Nadal and Parga, 1994), is known as the output equalization rule because the output $h^*(s)$ is uniformly distributed within its range limit. We will informally refer to both "L_0-min" and the infomax solution in the remainder of this paper. When $p = 2$, the optimal solution in Eq. (4.26) minimizes the mean square error of the reconstructed stimulus. This solutions was first proposed for optimal RGB color perception (Twer and MacLeod, 2001) and discussed in (Wang et al., 2012).

To summarize, the solution in Eq. (4.23) provides a systematic understanding of the optimal nonlinearities for the various criteria parametrized by p. In Figure 9 we illustrate different L_p optimal tuning curves for a standard Gaussian stimulus prior. Intuitively, the efficient coding problem can be understood as optimizing the allocation of neural descriptive power across an inhomogeneous stimulus distribution. Depending upon the value of p, the optimal allocation strategy balances between more frequently appearing stimuli with less frequent ones. Strategies corresponding more with Infomax (p near 0) emphasize stimuli with higher likelihood of appearing. On the other hand, L_p-optimal strategies with large p are more conservative and need to spend more resources to encode more surprising stimuli since the error penalty is larger.

4.2.4. Examples of Various Stimulus Prior Distributions

We applied our framework to various stimulus distributions (priors). For simplicity we only show the optimal tuning curve under the constant Gaussian noise assumption. In particular, we considered prior distributions that follow the generalized Gaussian model with scale parameter c and shape parameter β. From Eq. (4.23), the L_p-optimal tuning

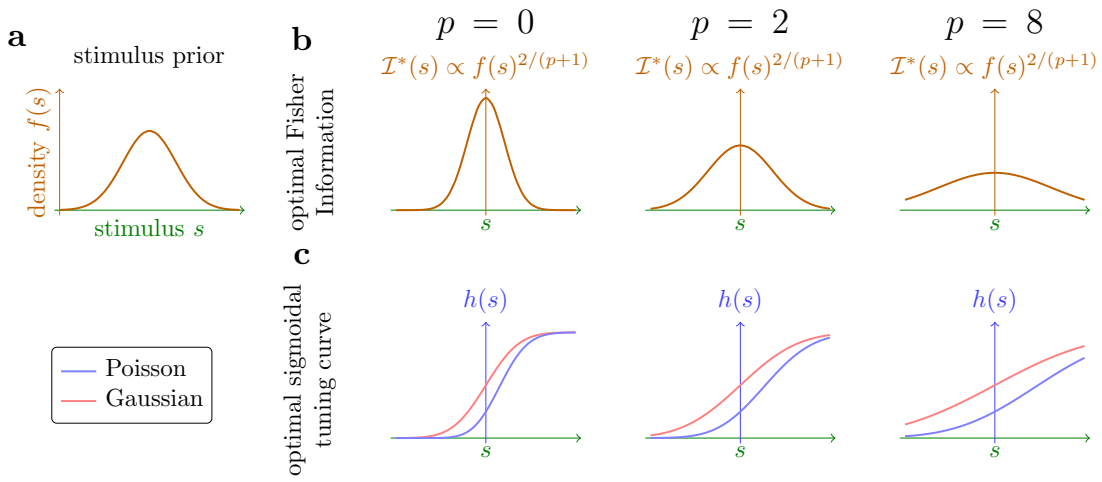

Figure 9: The L_p optimal sigmoidal tuning curves for for $p = 0, 2, 8$ for both Poisson or constant Gaussian noise models. **(a)** the Gaussian stimulus distribution (prior). **(b)** for each p, the optimal Fisher Information $\mathcal{I}^*(s)$ is derived based on the prior distribution **(c)** The optimal tuning curve for Poisson noise (blue lines) or constant Gaussian noise (red lines).

curve is related to the input stimulus distribution:

$$f(s) \propto \exp\left(-c|s|^\beta\right) \tag{4.27}$$

$$h'(s) \propto f(s)^{1/(1+p)} \propto \exp\left(-c\left(\frac{|s|}{(1+p)^{1/\beta}}\right)^\beta\right). \tag{4.28}$$

Therefore for a certain value of p, the nonlinearity is simply a rescaled version of the cumulative function of $f(s)$. The scalar $(1+p)^{1/\beta}$ is a decreasing function of β. In Figure 10 we illustrate three different cases: in the extreme of uniform distribution case where $\beta = \infty$, the scalar remains a constant and there is no difference across all the L_p-optimal tuning curve; for the Gaussian distribution case where $\beta = 2$, the scalar grows sub-linearly as $(1+p)^{1/2}$; for the Laplacian distribution case where $\beta = 1$, the scalar grows linearly as $(1+p)$.

Another important conclusion we would like to highlight is that all the L_p-optimal solutions except L_0 are not invariant under nonlinear stimulus transforms. For example, the L_2-optimal solution for a positive valued stimulus is not the same as the L_2-optimal solution for

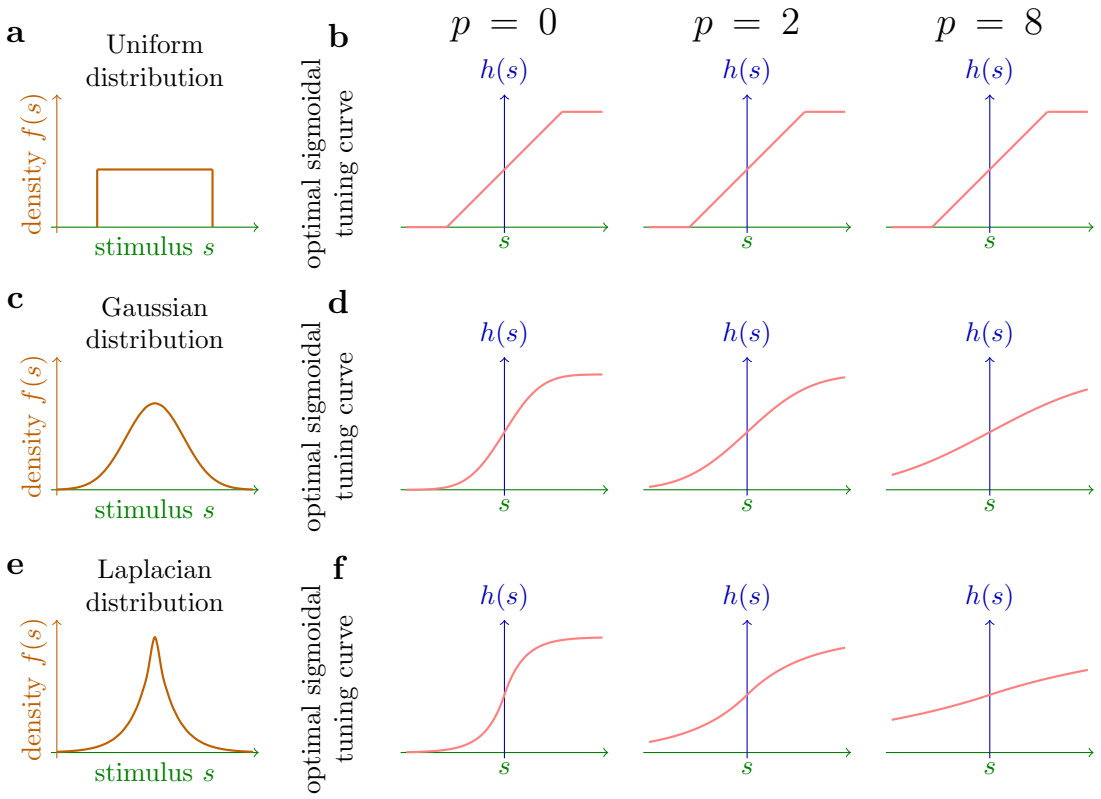

Figure 10: The L_p optimal sigmoidal tuning curve of a single neuron with constant Gaussian noise model. Here we compare the results for various form of prior distributions: uniform distribution **(a)**-**(b)**, Gaussian distirbution **(c)**-**(d)** and Laplacian (or double exponential) distribution **(e)**-**(f)**.

the log-stimulus. The L_0 is the only solution that is invariant under any one-to-one stimulus transformations. This fact again demonstrates the intuition that L_p-min strategies are highly task-driven – the solution changes if the stimulus variable undergoes some nonlinear transformation before being processed.

4.3. Generalization to Neural Populations

Here we show how to generalize the result of a single neuron to a neural population. The optimal neural population has been extensively studied (Zhang and Sejnowski, 1999; Pouget et al., 1999; Kang et al., 2004; McDonnell and Stocks, 2008; Nikitin et al., 2009; Ganguli and Simoncelli, 2010; Yaeli and Meir, 2010). The conclusions from these studies largely depend on the assumptions about the population being made.

4.3.1. Neural Population Assumptions

We also need to make certain restrictive assumptions. Rather than allowing all neurons in the population to independently exhibit arbitrary nonlinear tuning curves, we limit the tuning curve of the k-th neuron to have the following form

$$h_k(s) = h_0(\psi(s) - \psi(s_k)) \tag{4.29}$$

where $\psi(s)$ is the *meta-tuning curve* which transforms the stimulus s. For each neuron, s_k is the characteristic stimulus associated with that neuron. For example, s_k can be the preferred stimulus (at which the neuron elicits maximum neural response) for neurons with unimodal tuning curves or the semi-saturation stimulus (at which the neuron elicits half of maximum neural response) for neurons with sigmoidal tuning curves.

Below we denote $\tilde{s} = \psi(s)$ and $\tilde{s}_k = \psi(s_k)$ resulting from the output of the meta-tuning curve. We also assume the following:

(a) All neurons in the population share the same given nonlinearity $h_0(\tilde{s} - \tilde{s}_k)$.

(b) The characteristic stimuli $\tilde{s}_k$ are uniformly distributed, in other words the spacing $\Delta\tilde{s} = \tilde{s}_k - \tilde{s}_{k-1}$ between adjacent neurons is a constant.

(c) h_0 and h_0' are slowly varying when measured at the scale of $\Delta\tilde{s}$, i.e. $h_0(\tilde{s}_k) \approx h_0(\tilde{s}_k + \Delta\tilde{s})$ and $h_0'(\tilde{s}_k) \approx h_0'(\tilde{s}_k + \Delta\tilde{s})$. When $\Delta\tilde{s}$ is small, this constraint is equivalent to h_0 and h_0' being continuous.

(d) The neurons have independent output noise so the total Fisher information of the population is the linear sum of each individual ones $\mathcal{I}_{\text{total}}(s) = \sum_k \mathcal{I}_k(s)$ (see Section 2.2).

These assumptions are sometimes referred to as the "uniform tiling" properties of a neural population (Ganguli and Simoncelli, 2010; Grabska-Barwinska and Pillow, 2014). It is

important to note that the assumptions (a) and (b) limit the solutions to a sub-space of all possible population codes for which the mapped stimulus $\tilde{s}$ is encoded by a homogeneous population (see Figure 11). In our model, the total Fisher information of the population with either the Poisson noise or constant Gaussian noise (see Eq. (4.9) or Eq. (4.10)) becomes:

$$\mathcal{I}_0 \approx \mathcal{I}_{\text{total}}(\tilde{s}) = \sum_k \mathcal{I}_k(\tilde{s}) = \sum_k \frac{h_0'(\tilde{s} - \tilde{s}_k)^2}{h_0(\tilde{s} - \tilde{s}_k)} \quad \text{or} \quad \sum_k h_0'(\tilde{s} - \tilde{s}_k)^2 \tag{4.30}$$

The form of $h_0(\cdot)$ is fixed and often assumed but not limited to be either unimodal or sigmoidal. In Figure 11 we illustrate how to determine the individual tuning curves of the inhomogeneous neurons in the population.

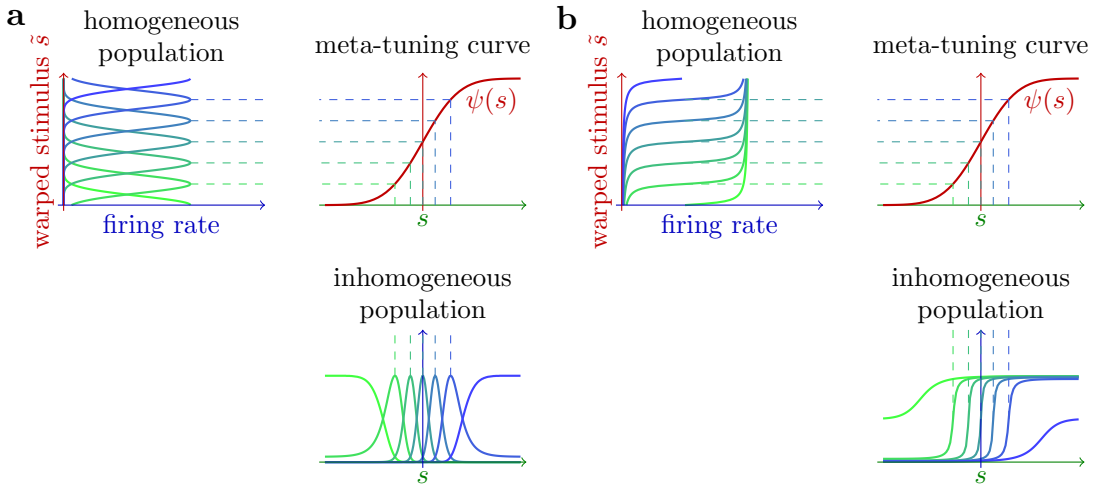

Figure 11: Under our assumptions, the inhomogeneous neural population tuning is derived by mapping a homogenous tuning description through the meta-tuning curve. via the sigmoidal meta-tuning curve $\psi(s)$. Two representative choices of h_0 are **(a)** unimodal and **(b)** sigmoidal.

4.3.2. Optimal Meta-tuning Curve

For any meta-tuning curve $\tilde{s} = \psi(s)$, we can calculate the Fisher Information of the k-th neuron and the total Fisher information for the population, with respect to the original

stimulus s as

$$\textbf{P:} \quad \mathcal{I}_k(s) \propto \frac{h_0'(\psi(s)-\tilde{s}_k)^2}{h_0(\psi(s)-\tilde{s}_k)} \cdot \psi'(s)^2 \tag{4.31}$$

$$\textbf{cG:} \quad \mathcal{I}_k(s) \propto h_0'(\psi(s)-\tilde{s}_k)^2 \cdot \psi'(s)^2 \tag{4.32}$$

$$\textbf{gG:} \quad \mathcal{I}_k(s) \propto \frac{h_0'(\psi(s)-\tilde{s}_k)^2}{h_0(\psi(s)-\tilde{s}_k)^\alpha} \cdot \psi'(s)^2 \tag{4.33}$$

$$\textbf{P, cG or gG:} \quad \mathcal{I}_{\text{total}}(s) = \sum_k \mathcal{I}_k(s) \approx \mathcal{I}_0 \cdot \psi'(s)^2 \tag{4.34}$$

In the population coding case, the mean L_p reconstruction error of s is related to the total Fisher information and we need to minimize the following term

$$\left\langle (\mathcal{I}_{\text{total}}(s))^{-p/2} \right\rangle_s = \int f(s)\,(\mathcal{I}_{\text{total}}(s))^{-p/2}\,ds \tag{4.35}$$

where $f(s)$ is the prior distribution of the stimulus s. We can limit the output of a non-decreasing meta-tuning curve to the range $0 \le \psi(s) \le const.$ Then minimizing the L_p reconstruction error is equivalent to the following optimization in terms of the meta-tuning curve $\psi(s)$:

$$\text{minimize} \qquad \left\langle (\mathcal{I}_{\text{total}}(s))^{-p/2} \right\rangle_s \approx \mathcal{I}_0^{-p/2} \cdot \int f(s)\psi'(s)^{-p}\,ds \tag{4.36}$$

$$\text{subject to} \qquad \int \psi'(s)\,ds \le const. \tag{4.37}$$

This optimization problem is the same as the constant Gaussian noise case we previously discussed in Section 4.2.3. This leads to a solution for the optimal meta-tuning curve $\psi^*(s)$ with corresponding total Fisher information:

$$\psi^{*\prime}(s) \propto f(s)^{1/(1+p)}, \quad \mathcal{I}^*_{\text{total}}(s) \propto f(s)^{2/(1+p)} \tag{4.38}$$

$$\psi^*(s) = \frac{\int_{-\infty}^{s} f(\xi)^{1/(1+p)}\,d\xi}{\int_{-\infty}^{\infty} f(\xi)^{1/(1+p)}\,d\xi} \tag{4.39}$$

This result illustrates that under our model, the Fisher Information allocation for the population is entirely determined by the meta-tuning curve $\psi(s)$, in the same fashion as the Fisher information allocation is determined by the sigmoidal tuning curve $h(s)$ of a single neuron with constant Gaussian noise. In Figure 12 we show the L_0, L_2 and L_8 optimal neural populations to encode a Gaussian stimulus random variable. Compared to previous work by Ganguli and Simoncelli (2010), our framework considers a more constrained class of neural populations because it assumes a fixed gain across neurons. Our formulation, however, allows us to specify an entire family of L_p-optimal solutions that smoothly incorporate the special cases of the Infomax and the MSE solutions.

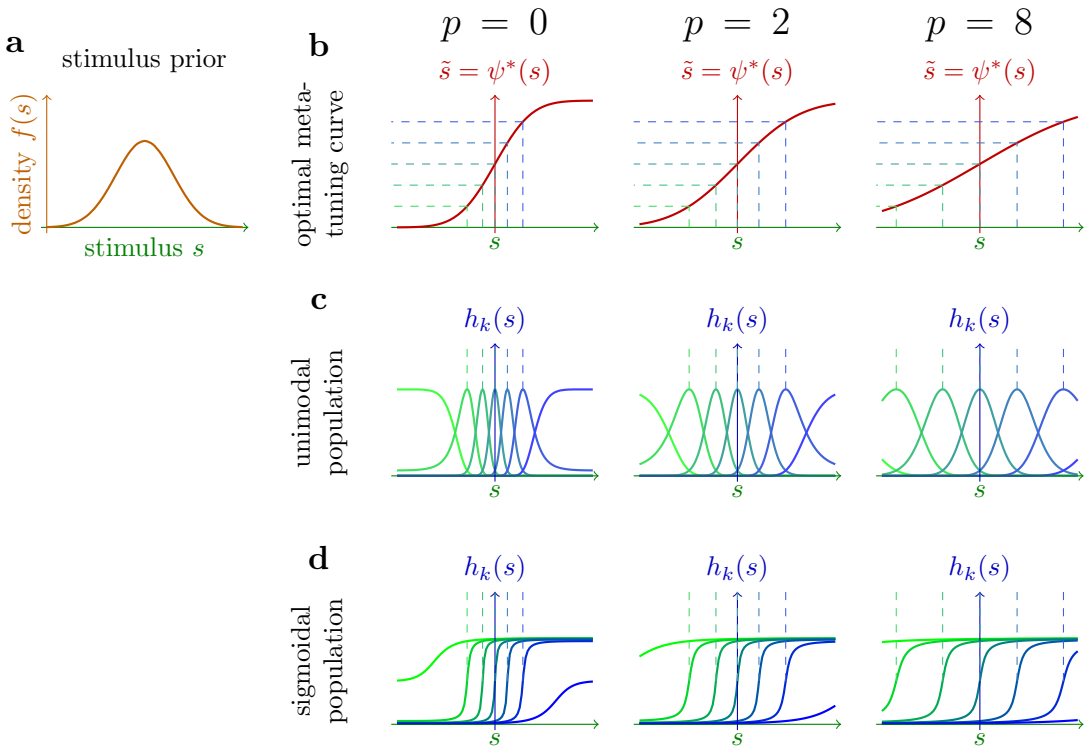

Figure 12: The L_p optimal neural populations for $p = 0, 2, 8$. Panels **(a)**, **(b)** are replicated from Figure 9 and the optimal meta-tuning curve for the population is identical to the optimal tuning curve of a neuron with constant Gaussian noise. Here we show two different kinds of optimal neural population, where each neuron has **(c)** unimodal tuning curves or **(d)** sigmoidal tuning curves.

4.4. Relaxing the Asymptotic Assumptions

For both the single neuron case and the neural population case, our results so far have relied on several key assumptions. The most restrictive one is the assumption that neurons are operating in the asymptotic long time limit. In this limit, the optimal decoder naturally converges to the maximum likelihood estimator. In contrast, in a more realistic scenario where encoding time is short, it is generally the case that a Bayesian (and usually biased) decoder will perform better. Unfortunately it is difficult to derive analytic solutions in this case yet numerical efforts have been made (Bethge et al., 2003; Nikitin et al., 2009). Furthermore, the derivation of the optimal Bayesian decoder can be intractable for arbitrary prior distributions.

In order to provide a sense of how well our derived analytic solutions hold for shorter encoding times, we compared their predicted performance to the actual measured performance obtained by numerical simulations. The decoding performance of our L_p optimized coding solutions can be easily simulated for arbitrary encoding times. For reasons of simplicity, we considered a standard Gaussian stimulus distribution $p(s)$ in our simulations. The encoding process is straightforward: stimuli are sampled and encoded by the L_p optimal code with additional Poisson spiking noise. For the decoding process, we examined both the assumed unbiased, maximum-likelihood estimator (MLE) and the maximum a posteriori estimator (MAPE). In both cases, iterative gradient descent method (Newton's method) was used to find the stimulus with maximal likelihood (for MLE) or maximal posterior likelihood (for MAPE). The mean L_p decoding error was then calculated over a large set of generated stimuli and compared to the theoretical prediction.

For a neuron with maximum firing rate $r_{\max}$ and a fixed length of the time window T, the key variable is the maximum allowed spike-count $N_{\max} = r_{\max}T$. For each value of $N_{\max}$ we run a total of 100 independent trials and in each trial, 100,000 stimuli were randomly generated. This experiment was done for both a single neuron with sigmoidal tuning curve and for a population of neurons with unimodal tuning curves. Results are shown in Figure 13. As

expected, the theoretical predictions were more accurate when $N_{\max}$ was large, with the critical value for $N_{\max}$ increasing as a function of p. For shorter encoding time, our result shows that the MAPE is a better estimator despite the similar performance for larger $N_{\max}$. The performance of the MLE seems to be lower bounded by our theoretical prediction (see the solid line) but the MAPE benefits from the prior information and is upper bounded by a constant related to that prior.

In the single neuron case, the critical spike-count $N_{\max}$ ranges from approximately 10^2 (for $p = 0.01$) to approximately 10^4 spikes (for $p = 2$). For some sensory neurons, such as the H1 neuron of a blowfly (see Section 5.1), the maximal firing rate $r_{\max}$ can be as high as 100Hz which means that the critical time for the long encoding assumption to be valid is around $T \geq 1$ sec (for $p = 0.01$) to $T \geq 100$ sec (for $p = 2$). In the neural population case, we run simulations with $K = 11$ neurons with unimodal tuning curves. As expected, the performance in terms of the L_p error is one order of magnitude better than for the single neuron case. Correspondingly, the critical spike-count $N_{\max}$ is much smaller: from approximately $10^{0.5}$ (for $p = 0.01$) to approximately $10^{1.5}$ spikes (for $p = 2$). For small p values, the performance matches the theoretical prediction for populations containing as few as 11 neurons with $N_{\max} \geq 3$ spikes per neuron. For larger p value such as $p = 2$, this number may increase to $N_{\max} \geq 30$ spikes per neuron.

In sum, we found that depending on the value of p the long time-limit assumptions can be reasonably relaxed for short encoding times. In particular, we find that the critical spike-count can be as low as $N_{\max} = 3 \sim 30$ spikes per neuron which justifies the biological relevance of our result. Generally, the predictions of our framework are much less constrained for smaller p values. We have also found that the performance of a Bayesian decoder (the MAPE) tends to be better than the MLE decoder, which shows that the optimality of our solution (MLE) strongly rely on the unbiased assumption. Fortunately, this limitation is subordinated to the short encoding time limitation. The MAPE itself is asymptotically unbiased and has similar performance as the MLE decoder once the critical

$N_{\max}$ is reached.

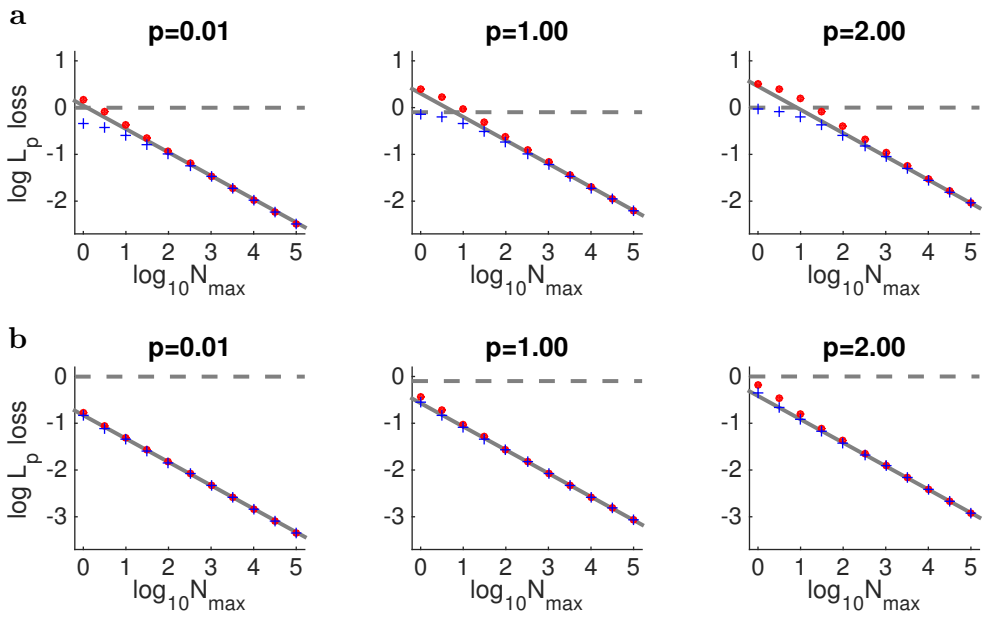

Figure 13: The simulated L_p encoding error (MLE: red dot, MAPE: blue cross) versus theoretical prediction assuming unbiased estimator (solid lines) or using only prior information (dashed lines). The markers indicates the median over 100 trials. **(a)** The performance of a single neuron with sigmoidal tuning curve (see e.g. Figure 10d). **(b)** The performance of a population with $K = 11$ neurons with unimodal tuning curves (see e.g. Figure 12c). The vertical axis is the mean L_p loss $\langle|\hat{s} - s|^p\rangle^{1/p}$ and the horizontal axis is $N_{\max}$, both in logarithm space with base 10.

4.5. Efficiency Criteria Used in Early Visual Perception Systems

Our theoretical analysis raises the question of which efficiency criterion the brain actually uses to encode information. In this section, we considered several different modalities in early visual perception: motion encoding, orientation encoding and contrast encoding. In each case, we attempted to estimate the prior distribution of the input stimulus and compared the tuning characteristics of the predicted efficient coding model with published physiological data.

4.5.1. Speed Perception by a Single Blowfly H1 Neuron

We first analyze data from the H1 neuron of blowfly, which encodes the speed s of a horizontally moving bar. The analyzed dataset (de Ruyter van Steveninck et al., 1997) was collected from a fly H1 neuron responding to a stochastically generated visual motion stimulus. The data was taken for 20 minutes at a sampling rate of 500Hz. For our purposes, we bin the dataset into 1200 bins with duration $\Delta t = 1$ second and we calculate the average stimulus s_i and the number of spikes N_i for $i = 1, \ldots, 1200$ and the stimulus-response relationship is plotted as dots in Figure 14a.

The natural speed prior for the blowfly is unknown. However, based on the investigation of natural movie clips, previous research has proposed that the prior distribution for visual speed should follow a power-law function of the form $f(s) \propto (1 + |s|/v_0)^{-2}$, where $v_0 > 0$ is a scale parameter (Van Hateren, 1993; Dong and Atick, 1995). For this particular form of the prior, the optimal L_p tuning curve $h_p^*(s)$ for a neuron with Poisson noise can be analytically computed.

$$h_p^{*\prime}(s) \propto f(s)^{\frac{1}{1+p}} \quad \Rightarrow \quad h_p^*(s) \propto \left(1 + \text{sign}(s)\left(1 - \frac{1}{(1+|s|/v_0)^{\frac{1-p}{1+p}}}\right)\right)^2 \tag{4.40}$$

It can be seen that for this parametric form of prior distribution, the L_p optimal solution exists only for $0 \leq p \leq 1$. In order to infer the prior distribution and the optimality criterion, we optimize the parameters v_0 and p to maximize the data likelihood. The result in Figure 14b shows the predicted speed prior distribution to which the H1 neuron is most likely adapted to. In Figure 14c-d we can see that $v_0 = 21.3$ deg/sec and $p = 0$ achieves the best data likelihood. However other pairs of (p, v_0) for $p < 0.8$ also yield good likelihood scores for this data.

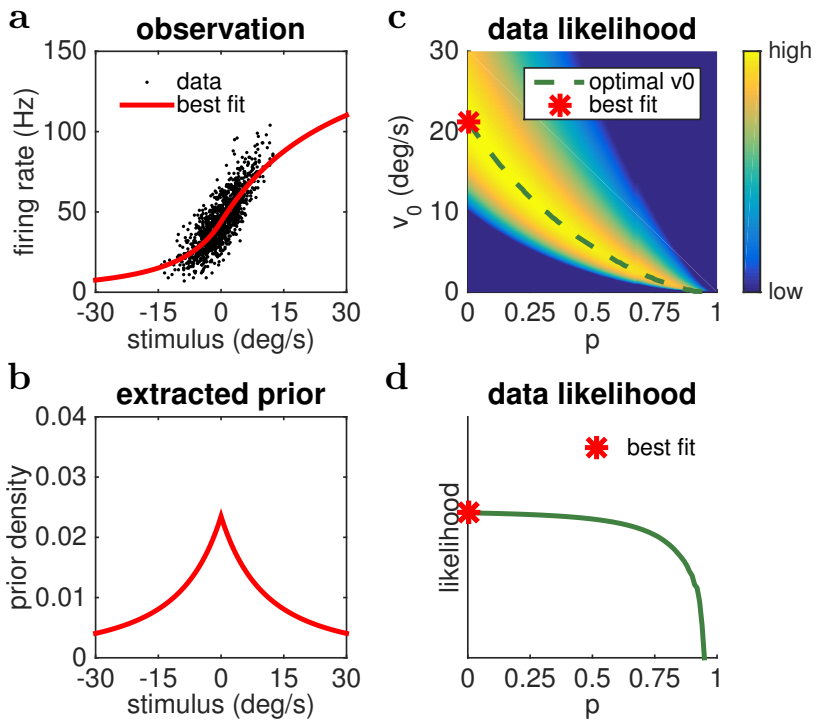

Figure 14: **(a)** the stimulus-response data collected from a fly H1 neuron (de Ruyter van Steveninck et al., 1997) and we plot the best tuning curve using the parametric model in Eq. (4.40). **(b)** the predicted prior distribution to which the fly H1 neuron is most likely adapted. **(c)** the optimal parameter v_0 and p is chosen to maximize the data likelihood. Dash line shows the optimal parameter $v_0(p)$ as a function p. **(d)** The maximum data likelihood for each pair $(p, v_0(p))$ as a function of p.

4.5.2. Population Code in Orientation Encoding

We also applied our proposed framework to analyze biological neural populations that encode local visual orientation. We first estimated the prior distribution $f(\theta)$ of local visual orientation θ from a natural image dataset (van Hateren and van der Schaaf, 1998) using a filter analysis at a single spatial scale (detailed description in Appendix 4.7). The resulting prior distribution is shown in Figure 15c and is very similar to previously estimated distributions (see *e.g.* Girshick et al. (2011)). Based on the estimated prior density, we derived the optimal meta-tuning curves $\psi(\theta)$ for various values of the norm parameter p (see Figure 15b). The unimodal tuning curves of the population (see Figure 15d) were then determined as described in Section 3.2 assuming an homogeneous population of certain

tuning width $\tilde{w}$ (see Figure 15a). Below we compare predictions of the model population with measured biophysical characteristics of orientation tuned neurons.

The first prediction is with regard to neural density. De Valois and colleagues reported that the ratio between neurons tuned for oblique *vs.* cardinal orientations is about 0.66 in area V1 of the macaque (Valois et al., 1982). In our framework the neural density as a function of θ is directly related to the derivative of the meta-tuning curves (Figure 15f). In order to compute the ratio between the number of neurons tuned for the oblique *vs.* the cardinal orientations, we binned the neural population into two sub-populations shown as blue/red regions in Figure 15f. The predicted ratio is a function of the norm parameter p (Figure 15e); for $p \approx 0.37$ the ratio of the model population matches the ratio found for neurons in V1.

We can also predict how the tuning width depends on the preferred stimulus of the neurons. Following the definition of Ringach et al. (2002), we defined the tuning width w as the length of the orientation interval over which a neuron's mean response is at least $1/\sqrt{2}$ of its peak firing rate. Figure 15h shows the predicted tuning width $w(\theta)$ as a function of the preferred orientation θ of a neuron in the model population. Each curve shows the tuning width $w(\theta)$ for a different assumed constant tuning width $\tilde{w}$ in the homogeneous population (Figure 15a). From these continuous functions we calculated the first and third quartiles w_{1Q}, w_{3Q} of the tuning widths across the inhomogeneous population. For each p value, the possible values of $w_{1Q}(\tilde{w})$ and $w_{3Q}(\tilde{w})$ form a curve with parameter $\tilde{w}$ as shown in Figure 15g. A comparison of the quartile predictions with physiological data from neurons in area V1 of the macaque (Ringach et al., 2002) suggests that the model best matches the data for a norm parameter of value $p = 0.08$.

Finally, we can make predictions about tuning curve asymmetries. Specifically, we compared the predicted asymmetry index (Henry et al., 1974) of our model population with the values found for biological neurons. Similar to the tuning width, the predicted asymmetry index is also a function of the assumed tuning width $\tilde{w}$ of the neurons in the homogeneous population

(see Figure 15j). We computed the predicted relationship between the mean asymmetry index and the median tuning width for different p value and compared it with measurements from simple cells in striate cortex of the cat (Henry et al., 1974). The reported median tuning width (measured at 1/2 peak amplitude; we have rectified our predictions accordingly) of 34° and asymmetry index 1.26 matches our predictions for $p \approx 0.85$ (see Figure 15i).

In summary, we found that the measured orientation tuning characteristics of neurons in primary visual cortex of the macaque and the cat match those model predictions that correspond to fairly low values of p.

4.5.3. Population Code in Contrast Perception

We also applied our framework to make predictions for the contrast gain characteristics of neurons in early visual cortex. The contrast of natural images has been defined in multiple ways in the literature. Two standard definitions of local contrast are the root-weighted-mean-square contrast (Najemnik and Geisler, 2005; Mante et al., 2005) and the equivalent-Michelson contrast (Brady and Field, 2000; Tadmor and Tolhurst, 2000; Clatworthy et al., 2003). We use the equivalent-Michelson contrast in order to match our predictions with recorded physiological data (Clatworthy et al., 2003). We gathered a total of 200,000 patches of size 32x32, randomly sampled from natural images from the dataset (van Hateren and van der Schaaf, 1998). The histogram of their equivalent-Michelson contrast is regarded as the prior distribution of the environment (see Figure 16c). The detailed description of this process is discussed Section 4.8.

In early visual perception systems, contrast information is encoded by a population of neurons with contrast selectivity in a soft-thresholding manner. One traditional model characterizes the neuron's response as a function of the contrast c via the Naka-Rushton equation (Naka and Rushton, 1966),

$$h(c) = h_{\max} \cdot \frac{c^q}{c_{50}^q + c^q} \tag{4.41}$$

where $h_{\max}$ is the maximum possible firing rate, c_{50} is the semi-saturation contrast so that $h(c_{50}) = 0.5 \cdot h_{\max}$ and q is an exponent parameter characterizing the steepness of the curve near c_{50}. Using our framework, we can predict the distribution of semi-saturation constant c_{50} within a population and compare that to physiology data (Clatworthy et al., 2003) (see Figure 16e). Our prediction suggests that the monkey V1 neurons are roughly performing infomax ($p \approx 0.15$) strategy while the cat striate cortex neurons are using a larger value of p ($p \approx 0.75$). As we can see from Figure 16e, the fit for c_{50} distribution of cat striate cortex is worse than the fit for c_{50} distribution of monkey's V1. The neural population in cat V1 seems to be adapted to smaller contrast values. This may be due to the mismatch between the natural image dataset and the true visual environment of the animal.

4.6. Discussion

In this paper we have proposed a family of efficiency criteria for neural coding. Each efficiency criterion uniquely determines an optimal way of encoding a scalar stimulus with an arbitrary prior distribution. The efficiency criteria are parametrized by a parameter $p \geq 0$ associated with the underlying goal of minimizing the L_p reconstruction error when using a maximum likelihood decoder. These efficiency criteria naturally generalize several special cases that have received much attention in the literature, e.g. the Infomax case ($p \to 0$) or the minimal mean squared error (MMSE) case ($p = 2$).

For each optimality criterion and a stimulus with known prior, we analytically derived the optimal tuning curve for a single neuron. To extend this result to determine optimal neural populations, we proposed to use the meta-tuning curve and showed that the optimal meta-tuning curve is identical to the optimal tuning curve for a single neuron with Gaussian noise. These predictions based upon different optimality criteria are tested against previously measured characteristics of several early visual systems for different animals. Predictions corresponding to low values of p provides the best match, which suggests that the optimality criterion is near Infomax for the neural representations being considered.

In our model and analysis, we have made the key assumption that the decoder is asymptotically unbiased. This implies that the results are strictly valid only in the low noise regime, *e.g.* when there is sufficient encoding time and/or a sufficient number of neurons. However, based on numerical simulations we found that it is reasonably safe to relax the long encoding time assumption in particular if the neural population size is large and/or the optimal criterion parameter p is small.

Many behavioral studies also suggest that human and other animals make decisions that are often biased due to the effects of prior beliefs (Knill and Richards, 1996; Wei and Stocker, 2015). With numerical simulations we showed that at short encoding times, the Bayesian MAPE decoder is indeed performing better than the unbiased MLE decoder, and slightly better than our analytic predictions. In fact, the performance of the MLE is lower bounded by our theoretical predictions (solid lines in Figure 13) while the performance of the MAPE benefits from the prior information. Thus our results are strictly valid only when assuming an MLE decoder.

In Section 4.2.2, we analyzed the Poisson noise model and the constant Gaussian noise model. Similar analysis can be applied to other noise models where the output variance depends upon the output mean. For neural populations, we assumed that the output noise of an individual neuron is independent from the others, thus simplifying the computation of the total Fisher information of the population. If the output noise has a correlated structure, then the total Fisher information is no longer the linear sum of the individual Fisher informations. Analysis of neural populations described by a meta-tuning curve with correlated noise is a subject for further investigation.

In conclusion, we believe that our model shows the utility of exploring different reconstruction error criterion for analyzing neural responses in perceptual systems. The parameter p describes whether the neural system is adapted to more or less robust error statistics, and we have obtained some estimates of this parameter from data on early visual processing neurons in a number of different animals. It will be interesting to explore how the parameter

p changes as information propagates through various stages of the perceptual system. We are also currently investigating how this analysis can be extended to higher-dimensional stimuli and to more complex noise models.

4.7. Appendix I: Estimating the Distribution over Local Orientation

We extracted orientation statistics for natural images from a standard image database (van Hateren and van der Schaaf, 1998). First we randomly sampled 200,000 square patches (16pix-by-16pix) across the entire database. We then created a set of sinewave grating filters with a fixed spatial frequency that was close to the human peak sensitivity (approximately 4 cycle per visual degree or 8 pixels/cycle) but various phase and 360 different orientations (0° to 179.5° with 0.5° spacing). The dominant orientation of each patch was determined by the maximum response across all these filters. To mitigate the effect of pixel-wise noise or quantization effects, we only used those patches with high filter response levels (top 50%). The resulting prior distribution is very similar to previously measured distributions (e.g. Girshick et al. (2011)) and is shown in Figure 15c. We used a spline function to fit the cumulative of the empirical histogram in order to obtain a smooth version of the density $f(\theta)$.

4.8. Appendix II: Equivalent-Michelson Contrast

Originally, the Michelson contrast is defined for sinusoid gratings based on its max/min luminance

$$c = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}} \tag{4.42}$$

It is clear the the Michelson contrast has a value between 0 and 1. For any patches of non-sinusoid gratings, we determine its equivalent-Michelson contrast in the following way.

For each image patch, we use a set of 64 odd-Gabor filters $g_{\text{gabor}}(x, y)$ of different orientation θ and wavelength λ to convolute with natural image patches to obtain local responses.

Specifically, the Gabor filters are

$$g_{\text{gabor}}(x,y) = g_{\text{normal}}(x,y) \cdot g_{\text{sinusoid}}(x,y) \tag{4.43}$$

$$g_{\text{normal}}(x,y) = \exp\left(-\frac{x'^2 + y'^2}{2\sigma^2}\right), \quad g_{\text{sinusoid}}(x,y) = \sin\left(2\pi\frac{x'}{\lambda}\right) \tag{4.44}$$

$$x' = x\cos\theta + y\sin\theta, \quad y' = -x\sin\theta + y\cos\theta, \quad \sigma = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\frac{2^b+1}{2^b-1}\lambda \tag{4.45}$$

where the orientation θ takes 8 values uniformly sampled from the range $[0, \pi]$, the wavelength λ takes 8 values uniformly sampled in the logarithm space from 4 to 85.3 pixels per cycle. The size of Gaussian filter σ is automatically determined by the wavelength λ and a fixed octave value $b = 1.5$ in order to best match the properties of simple cells in the primary visual cortex.

With such a filter bank of 64 Gabor filters, we calculate the equivalent-Michelson contrast for each image patches. For each Gabor filters, we use the corresponding Gaussian filters $g_{\text{normal}}(x,y)$ to compute the local mean luminance to model luminance adaptation. We also use the corresponding sinusoid filter $g_{\text{sinusoid}}(x,y)$ to construct a testing sinusoid grating $L_{\text{ave}} + L_{\text{amp}} \cdot g_{\text{sinusoid}}(x,y)$. By properly choosing the parameters L_{ave} and L_{amp}, we can match both the Gabor-filter response and the Gaussian-filter response. The equivalent-Michelson contrast is then determined by the Michelson contrast of this testing grating:

$$L_{\max} = L_{\text{ave}} + |L_{\text{amp}}|, \quad L_{\min} = L_{\text{ave}} - |L_{\text{amp}}| \quad \Rightarrow \quad c = \frac{|L_{\text{amp}}|}{L_{\text{ave}}} \tag{4.46}$$

The above process is summarized in Figure 17. The local contrast value of each image patches is then determined by taking the maximum among the 64 equivalent-Michelson contrast values calculated using the Gabor filter bank. This max operation is taken in order to match the normalization computation taken place in the visual perception pathway (Carandini and Heeger, 2012). Neurons that are responding to a low contrast value often appear to be silent (normalized out) when there is a neighbor neuron responding to a significantly larger contrast.

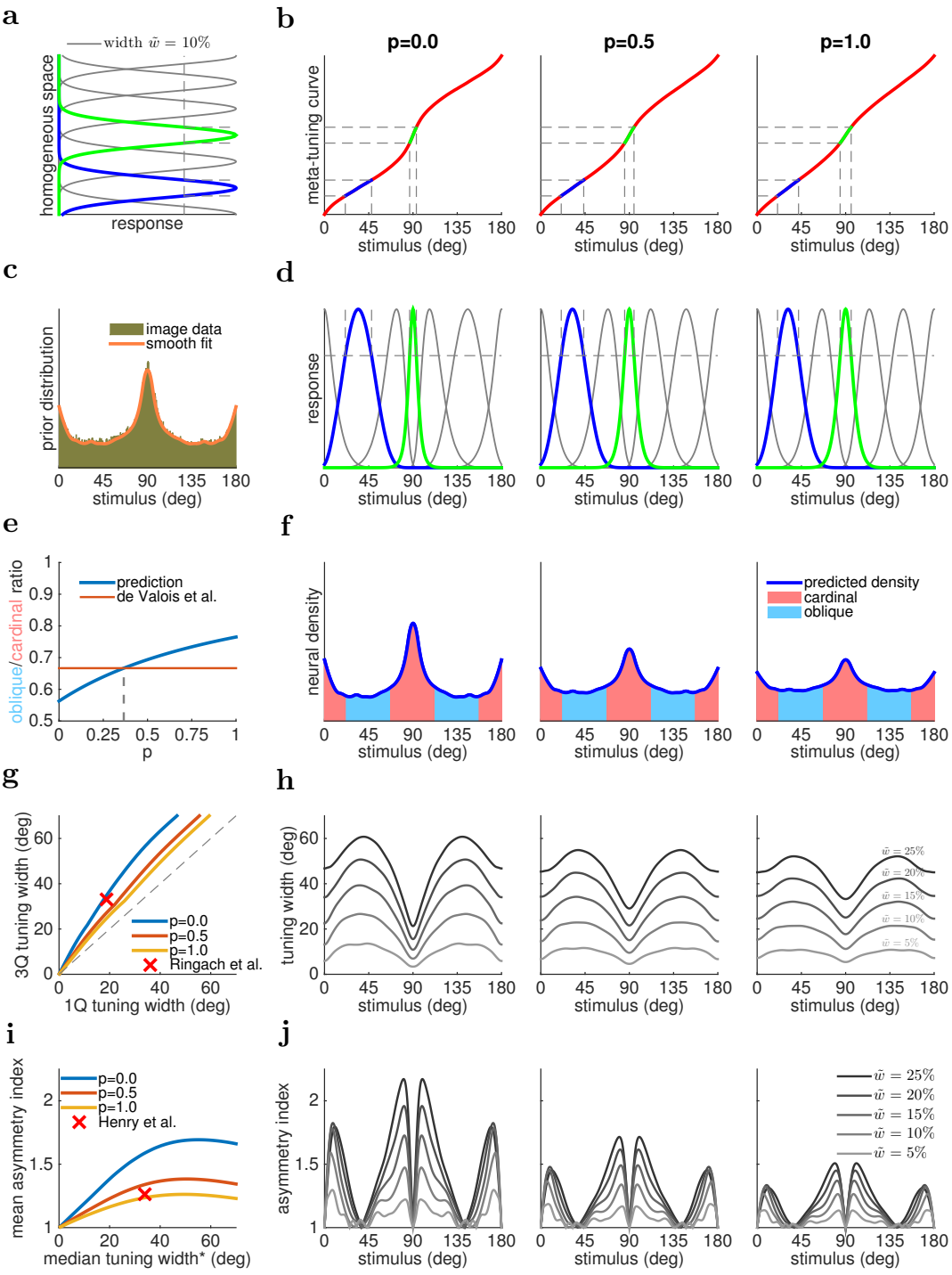
a
width $\tilde{w} = 10\%$
homogeneous space
response
b
p=0.0
p=0.5
p=1.0
meta-tuning curve
stimulus (deg)
c
image data
smooth fit
prior distribution
stimulus (deg)
d
response
stimulus (deg)
e
prediction
de Valois et al.
oblique/cardinal ratio
p
f
predicted density
cardinal
oblique
neural density
stimulus (deg)
g
p=0.0
p=0.5
p=1.0
Ringach et al.
3Q tuning width (deg)
1Q tuning width (deg)
h
tuning width (deg)
stimulus (deg)
$\tilde{w} = 25\%$
$\tilde{w} = 20\%$
$\tilde{w} = 15\%$
$\tilde{w} = 10\%$
$\tilde{w} = 5\%$
i
p=0.0
p=0.5
p=1.0
Henry et al.
mean asymmetry index
median tuning width* (deg)
j
asymmetry index
stimulus (deg)
$\tilde{w} = 25\%$
$\tilde{w} = 20\%$
$\tilde{w} = 15\%$
$\tilde{w} = 10\%$
$\tilde{w} = 5\%$

Figure 15: Comparison between theoretically predicted and physiologically measured tuning characteristics of orientation tuned neural populations. **(a)-(d)** cartoon examples of L_p-optimal neural population derived based on a homogeneous neural population and the optimal meta-tuning curve, which is determined by the prior distribution extracted from natural images. The p values are 0, 0.5 and 1. **(e)-(f)** the oblique versus cardinal ratio prediction is compared with previous results (Valois et al., 1982) on macaque V1 foveal neurons, which suggests $p \approx 0.37$. **(g)-(h)** the 1st and 3rd quartile tuning width prediction is compared with previous results (Ringach et al., 2002) on macaque V1, which suggests $p \approx 0.08$. **(i)-(j)** the asymmetry index and median tuning width(*) prediction is compared with previous results (Henry et al., 1974) on cat's striate cortex, which suggests $p \approx 0.85$. (* the tuning width here is measured at half amplitude to be consistent with previous study.)

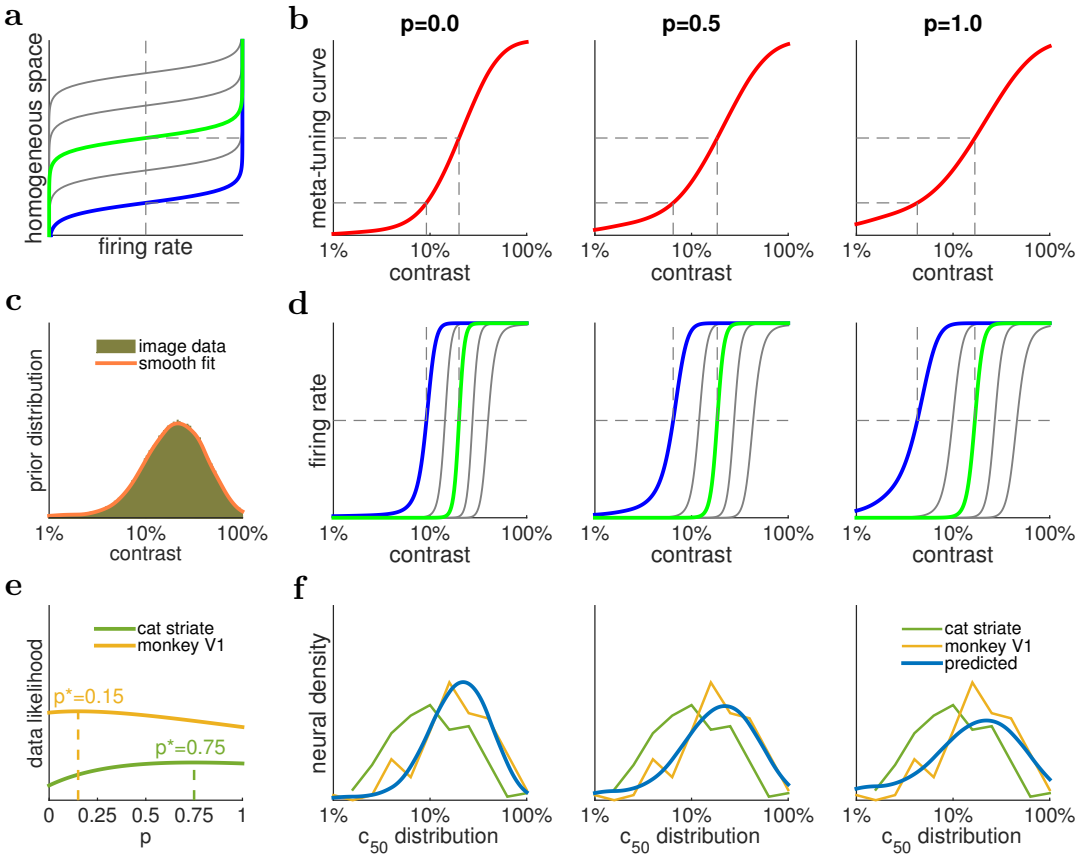

Figure 16: The analysis of optimal L_p optimal neural population to encode contrast value in natural images. **(a)-(d)** cartoon examples of L_p-optimal neural population are derived based on a homogeneous neural population and the optimal meta-tuning curve, which is determined by the prior distribution of equivalent-Michelson contrast extracted from natural images. The p values are 0,1,2. **(e)-(f)** the predicted of c_{50} distribution for the entire population is compared with physiology data reproduced from (Clatworthy et al., 2003) on cat's striate cortex and monkey's V1, which suggests $p \approx 0.15$ for the monkey and $p \approx 0.75$ for the cat.

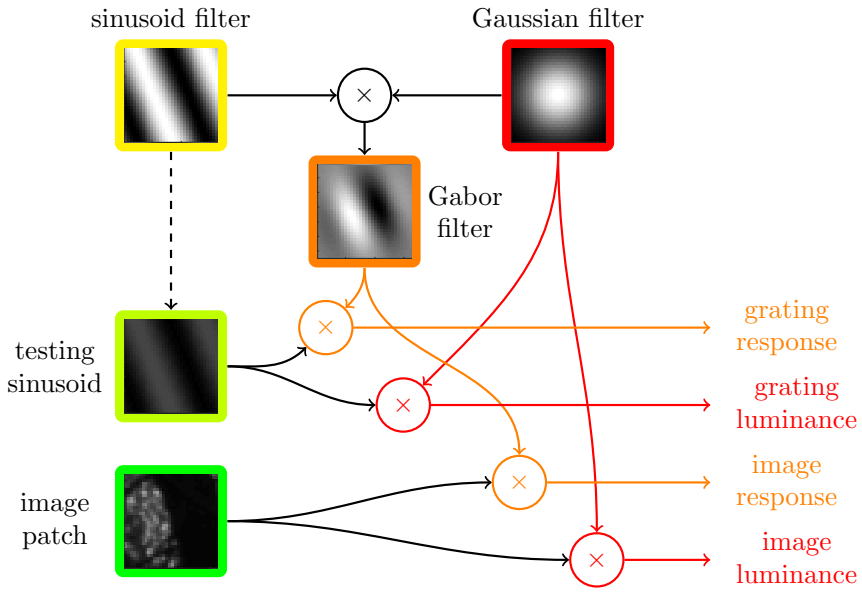

Figure 17: The process to determine equivalent-Michelson contrast for an image patch with respect to certain Gabor filter.

CHAPTER 5 : L_p Optimal Codes for High Dimensional Stimulus

5.1. Introduction

In this section we aim to further generalize results from Chapter 4 to multivariate stimulus $\mathbf{s}$. Here we present the result for the general L_p optimal criteria, non-Gaussian stimulus prior and possibly an over-complete neural population. As a special case, the optimal L_2 complete population for high-dimensional Gaussian stimulus has been discussed in our previous paper (Wang et al., 2013).

First we need to extend some notions to their high dimensional analogies because both the stimulus $\mathbf{s}$ and the its estimator $\hat{\mathbf{s}}$ are n-dimensional. We need to consider the high dimensional L_p-error of an n-dimensional vector $\hat{\mathbf{s}} - \mathbf{s}$.

$$\|\hat{\mathbf{s}} - \mathbf{s}\|_p = \left(\sum_{i=1}^{n} |\hat{s}_i - s_i|^p \right)^{1/p}. \tag{5.1}$$

This notion defines a norm only if $p \geq 1$ and a semi-norm if $0 < p < 1$. Now we wish to optimize the overall L_p loss under a similar set of constraints on the population

$$\text{minimize} \qquad \langle \|\hat{\mathbf{s}} - \mathbf{s}\|^p \rangle_{\boldsymbol{\epsilon},\mathbf{s}} \tag{5.2}$$

$$\text{subject to} \qquad \text{rank}(\mathbf{W}_{n\times m}) = n, \quad \sum_{k=1}^{m} g_k \leq g_{\text{total}} \tag{5.3}$$

$$0 \leq h_k(t) \leq 1 \text{ for } k = 1, \ldots, m. \tag{5.4}$$

Here the filter $\mathbf{W}$ is assumed to be full rank so it must be either complete or over-complete. The total gain of the population is given by g_{total} which describes a constraint which limits the output range of each neuron.

5.1.1. Objective Functions in terms of Fisher Information

The idea of Fisher information can also be extended to its matrix form for multivariate case. For each location of $\mathbf{s}$, the k-th neuron contributes a rank one matrix to the overall Fisher information matrix

$$\mathcal{I}_k(\mathbf{s})_{n\times n} = \left\langle \nabla_{\mathbf{s}} \log p(r_k|\mathbf{s}) \cdot \nabla_{\mathbf{s}} \log p(r_k|\mathbf{s})^T \right\rangle_{r_k} \tag{5.5}$$

The total Fisher information is the linear sum of these rank one matrices $\mathcal{I}(\mathbf{s}) = \sum_{k=1}^{m} \mathcal{I}_k(\mathbf{s})$ and it still holds true that the error vector $\hat{\mathbf{s}}-\mathbf{s}$ is asymptotically a Gaussian random variable with mean $\mathbf{0}$ and covariance matrix $\mathcal{I}(\mathbf{s})^{-1}$.

$$\hat{\mathbf{s}} - \mathbf{s} \sim \text{Normal}(\mathbf{0}, \mathcal{I}(\mathbf{s})^{-1}) \tag{5.6}$$

However, the L_p loss for vectors defined in Eq. (5.1) does not have a simple relationship with the Fisher information matrix except when $p = 2$. This is because the L_p error depends on the choice of the coordinate system and is not rotationally invariant in general. To resolve this issue, we choose a different way to define the L_p error in high dimensional spaces by using the eigenvalues of the Fisher information matrix, which is coordinate system free. Let us denote $\mathcal{I}(\mathbf{s}) = \mathbf{U}(\mathbf{s})^T \mathbf{\Lambda}(\mathbf{s}) \mathbf{U}(\mathbf{s})$ as the eigenvalue decomposition of $\mathcal{I}(\mathbf{s})$. We also denote $\mathbf{\Lambda}(\mathbf{s}) = \text{diag}(\lambda_1(\mathbf{s}), \ldots, \lambda_n(\mathbf{s}))$. For an asymptotically Gaussian random variable we can show that

$$\|\hat{\mathbf{s}} - \mathbf{s}\|_p^p \approx \text{tr}\left[\mathcal{I}(\mathbf{s})^{-p/2}\right] = \text{tr}\left[\mathbf{\Lambda}(\mathbf{s})^{-p/2}\right] = \sum_{i=1}^{n} \lambda_i(\mathbf{s})^{-p/2} \tag{5.7}$$

Using this result, one can show that the high-dimensional L_p-min and infomax problem is just Eq. (5.8) (see Section 2.3 for detailed derivation).

$$\text{minimize} \qquad \left\langle \text{tr}\left[\mathcal{I}(\mathbf{s})^{-p/2}\right] \right\rangle_s \tag{5.8}$$

Similar to the 1D case, p parametrically connects various criteria to measure the neural coding quality. As a special example, the above problem is equivalent to the minimum mean squared error (MMSE) problem when $p = 2$

$$\text{minimize} \qquad \left\langle \text{tr}\left[\mathcal{I}(\mathbf{s})^{-1}\right]\right\rangle \tag{5.9}$$

In the limit of $p \to 0^+$, denote $\mathbf{M} = \mathcal{I}(\mathbf{s})^{-1}$ and one can use matrix exponential of a positive-definite matrix

$$\mathbf{M}^{p/2} = \exp\left(\frac{p}{2}\log\mathbf{M}\right) = \mathbf{I} + \frac{p}{2}\log\mathbf{M} + O(p^2) \tag{5.10}$$

$$\Rightarrow \quad \text{tr}\left[\mathbf{M}^{p/2}\right] = \text{tr}\left[\mathbf{I}\right] + \frac{p}{2}\,\text{tr}\left[\log\mathbf{M}\right] + O(p^2) = n + \frac{p}{2}\log\det\mathbf{M} + O(p^2) \tag{5.11}$$

As p goes to zero, the leading order optimization problem is equivalent to the infomax problem (compare to Eq. (4.6) for the 1D case; see (Brunel and Nadal, 1998) for derivation of multivariate case)

$$\text{minimize} \qquad -\frac{1}{2}\left\langle\log\det\mathcal{I}(\mathbf{s})\right\rangle_s = -\frac{1}{2}\left\langle\text{tr}\left[\log\mathcal{I}(\mathbf{s})\right]\right\rangle \tag{5.12}$$

5.1.2. Constraints in terms of Fisher Information

The optimal nonlinearities of a neuron with Poisson noise, constant Gaussian noise or generalized Gaussian noise (see Eq. (4.12) - Eq. (4.14)) are equal to each other after raising to a proper power. For example, the optimal nonlinearities for Poisson neurons can be exactly derived by applying the square operation on optimal nonlinearities for Gaussian neurons. For the purpose of clarity, we will focus on the constant Gaussian noise case for the rest of the paper.

For a population of neuron with constant Gaussian noise, the individual Fisher information for each neuron and the total Fisher information for the population are given by (see

Section 2.2 for detailed derivation)

$$\mathcal{I}_k(\mathbf{s}) = g_k^2 \cdot h_k'(\mathbf{w}_k^T\mathbf{s})^2 \cdot \mathbf{w}_k\mathbf{w}_k^T \tag{5.13}$$

$$\mathcal{I}(\mathbf{s}) = \sum_{k=1}^{m} \mathcal{I}_k(\mathbf{s}) = \mathbf{W}\mathbf{G}\mathbf{H}(\mathbf{s})^2\mathbf{G}\mathbf{W}^T \tag{5.14}$$

where $\mathbf{W}$ is the linear filter, $\mathbf{G}$ and $\mathbf{H}(\mathbf{s})$ are diagonal matrices indicating the gain and the sensitivity at $\mathbf{s}$ for the population

$$\mathbf{W}_{n\times m} = \Big(\mathbf{w}_1, \ldots, \mathbf{w}_m\Big) \tag{5.15}$$

$$\mathbf{G}_{m\times m} = \begin{pmatrix} g_1 & & 0 \\ & \ddots & \\ 0 & & g_m \end{pmatrix} \tag{5.16}$$

$$\mathbf{H}(\mathbf{s})_{m\times m} = \begin{pmatrix} h_1'(\mathbf{w}_1^T\mathbf{s}) & & 0 \\ & \ddots & \\ 0 & & h_m'(\mathbf{w}_1^T\mathbf{s}) \end{pmatrix} \tag{5.17}$$

5.1.3. Full Optimization Problem and Its Variant

As we have discussed above, the objective function for L_p error minimization is

$$\text{minimize} \qquad \Big\langle \operatorname{tr}\Big[(\mathcal{I}(\mathbf{s}))^{-p/2}\Big]\Big\rangle = \Big\langle \operatorname{tr}\Big[\big(\mathbf{W}\mathbf{G}\mathbf{H}(\mathbf{s})^2\mathbf{G}\mathbf{W}^T\big)^{-p/2}\Big]\Big\rangle \tag{5.18}$$

$$\text{subject to} \qquad \operatorname{rank}(\mathbf{W}) = n, \quad \operatorname{tr}\big[\mathbf{G}^2\big] \le g_{\text{total}} \tag{5.19}$$

$$0 \le h_k(\cdot) \le 1, \quad k = 1, \ldots, m, \tag{5.20}$$

This problem can be analytically solved for special cases when $p = 2$ or the limit of $p \to 0$. For general value of p, this optimization problem is intractable because of the nonlinear entanglement between the expectation and the fractional matrix power in Eq. (5.18).

To resolve this issue, we consider an alternative form of the optimization problem. Instead

of assuming the best possible decoder $\mathbf{P}^*(\mathbf{s})$, we assume a fixed, unbiased decoder $\hat{\mathbf{s}} = \mathbf{P}^T\hat{\mathbf{t}}$ where $\mathbf{P}$ is one particular right-inverse of $\mathbf{W}$ (there could be many) which satisfies $\mathbf{WP} = \mathbf{I}_n$ and $\hat{t}_k = h_k^{-1}(g_k^{-1}r_k)$. We measure the asymptotic L_p loss of such decoder $\hat{\mathbf{s}}$ and the best possible $\mathbf{P}^*$ should attain the lower bounds provided by the Fisher information

$$\min_{\mathbf{P}} \left\langle \mathrm{tr}\left[\left(\mathbf{P}^T\mathbf{G}^{-1}\mathbf{H}(\mathbf{s})^{-2}\mathbf{G}^{-1}\mathbf{P}\right)^{p/2}\right]\right\rangle = \left\langle \mathrm{tr}\left[\left(\mathbf{WGH}(\mathbf{s})^2\mathbf{GW}^T\right)^{-p/2}\right]\right\rangle \tag{5.21}$$

$$\mathbf{P}^*(\mathbf{s}) = \mathbf{GH}(\mathbf{s})^2\mathbf{GW}^T\left(\mathbf{WGH}(\mathbf{s})^2\mathbf{GW}^T\right)^{-1} \tag{5.22}$$

As we have shown above, the optimal $\mathbf{P}^*(\mathbf{s})$ is a function of $\mathbf{s}$, which makes the problem intractable. However the optimal $\mathbf{P}^*(\mathbf{s})$ may become trivial under two circumstances: (1) if $\mathbf{GH}(\mathbf{s})^2\mathbf{G} = \lambda\mathbf{I}$ is a constant matrix or (2) if $\mathbf{W}$ is invertible.

In the first case, we will show that the optimal $\mathbf{P}^*$ is the pseudo inverse $\mathbf{W}^T(\mathbf{WW}^T)^{-1}$ and more discussion will be presented in Section 5.2. Under the second condition, the matrix $\mathbf{W}$ is $n \times n$ which indicates that the population is complete and $\mathbf{P}^*$ can be reduced to the ordinary matrix inverse $\mathbf{W}^{-1}$. This case is dealt with in Section 5.3. When neither of these two conditions is satisfied, we have to sacrifice the optimality of $\mathbf{P}$ if we want to solve the problem analytically. In particular, we fix a reasonable $\mathbf{P}$ and optimize the left side of Eq. (5.21) instead of the true objective function in Eq. (5.18).

5.2. Results for Linear Neurons

In this section, we consider a closely related but very important variant of the original problem. Instead of nonlinear transfer functions $g_k h_k(\mathbf{w}_k^T\mathbf{s})$, we assume the activation function is linear $h_k(x) = h_k \cdot x$. Under such assumption, all neurons are linear and can generate real valued outputs, the multiplicative factor g_k and h_k can be omitted because they can be represented in the linear projection $\mathbf{w}_k$. As a consequence, each linear neuron in the population simply calculates the linear projections $\mathbf{w}_k^T\mathbf{s}$ of the original variable and is subject to a constant Gaussian noise with equal variance $\boldsymbol{\epsilon} \sim \mathrm{Normal}(\mathbf{0}, \mathbf{I})$. In particular,

the response of each neuron in the population is given by

$$r_k = \mathbf{w}_k^T \mathbf{s} + \epsilon_k \quad \text{or matrix form} \quad \mathbf{r} = \mathbf{W}^T \mathbf{s} + \boldsymbol{\epsilon} \tag{5.23}$$

where $\mathbf{W}$ is a full rank projection matrix to be optimized. Let $\mathbf{P} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$ be the pseudo inverse matrix. In this case, the unbiased estimator in Eq. (5.22) is simply $\hat{\mathbf{s}} = \mathbf{P}^T\mathbf{r}$. The error vector $\hat{\mathbf{s}} - \mathbf{s} = \mathbf{P}^T\boldsymbol{\epsilon}$ is a Gaussian random variable with covariance $\mathbf{cov}(\hat{\mathbf{s}} - \mathbf{s}) = \mathbf{P}^T\mathbf{P}$ because the noise $\boldsymbol{\epsilon}$ is a standard Gaussian. It is clear that the Fisher information matrix given in Eq. (5.14) is indeed the inverse of the constant covariance matrix $\mathcal{I}(\mathbf{s}) = \mathbf{W}\mathbf{W}^T = (\mathbf{P}^T\mathbf{P})^{-1}$ for this projection matrix $\mathbf{P}$.

For arbitrary prior distribution $f(\mathbf{s})$, the decoding error $\hat{\mathbf{s}} - \mathbf{s}$ is independent of $\mathbf{s}$ and has identical Gaussian distribution no matter what the value of $\mathbf{s}$ is. Here, as usual, we want to minimize the L_p loss of this decoding error by choosing the optimal filters $\mathbf{W}$. Because there is no saturation constraints on $h_k(\mathbf{s})$ anymore, here we optimize the problem under the total power constraint, which assumes the total variance of all neuronal channels cannot exceed a certain amount

$$\text{total power} = \text{tr}\left[\mathbf{cov}(\mathbf{r})\right] = \text{tr}\left[\mathbf{W}^T\mathbf{C}\mathbf{W} + \mathbf{I}\right] = const. \tag{5.24}$$

where $\mathbf{C} = \mathbf{cov}(\mathbf{s})$ is the covariance of the stimulus variable. The objective function depends only on the linear filter $\mathbf{W}$

$$\underset{\mathbf{W}}{\text{minimize}} \qquad \text{tr}\left[\mathcal{I}(\mathbf{s})^{-p/2}\right] = \text{tr}\left[(\mathbf{W}\mathbf{W}^T)^{-p/2}\right] \tag{5.25}$$

$$\text{subject to} \qquad \text{tr}\left[\mathbf{W}^T\mathbf{C}\mathbf{W}\right] = c_{\text{total}}. \tag{5.26}$$

To solve the above problem, one can use the Singular Value Decomposition (SVD) of $\mathbf{W}_{n\times m} = \mathbf{U}_{n\times n}\mathbf{D}_{n\times m}\mathbf{V}_{m\times m}$ and optimize these matrices. In the SVD, $\mathbf{U}, \mathbf{V}$ are unitary matrices and $\mathbf{D}$ is a rectangular matrix with singular values $(d_1, \ldots, d_n)$ along the diagonal

and zero at all other off-diagonal entries. With some calculation one can show that

$$(\mathbf{W}\mathbf{W}^T)^{-p/2} = (\mathbf{U}\mathbf{D}\mathbf{D}^T\mathbf{U}^T)^{-p/2} = \mathbf{U}(\mathbf{D}\mathbf{D}^T)^{-p/2}\mathbf{U}^T \tag{5.27}$$

$$\mathbf{W}^T\mathbf{C}\mathbf{W} = \mathbf{V}^T\mathbf{D}^T\mathbf{U}^T\mathbf{C}\mathbf{U}\mathbf{D}\mathbf{V} \tag{5.28}$$

We can take the trace, rearrange the terms and denote $z_i = (\mathbf{U}^T\mathbf{C}\mathbf{U})_{ii}$. This would lead to an equivalent optimization problem

$$\underset{\mathbf{U},\mathbf{D},\mathbf{V}}{\text{minimize}} \qquad \text{tr}\left[(\mathbf{D}\mathbf{D}^T)^{-p/2}\right] = \sum_{i=1}^{n}(d_i^2)^{-p/2} \tag{5.29}$$

$$\text{subject to} \qquad \text{tr}\left[(\mathbf{D}\mathbf{D}^T)\cdot\mathbf{U}^T\mathbf{C}\mathbf{U}\right] = \sum_{i=1}^{n} d_i^2 z_i = c_{\text{total}}. \tag{5.30}$$

The optimization problem is now free of $\mathbf{V}$ therefore the optimal $\mathbf{V}$ can be any unitary matrix. On the other hand, for any fixed unitary matrix $\mathbf{U}$ (and fixed z_i), the optimal condition for d_i^2 is

$$(d_i^2)^{-p/2-1} - \lambda z_i = 0 \quad \Rightarrow \quad d_i^2 = \lambda_0 \cdot z_i^{-2/(p+2)} \tag{5.31}$$

If we plug this into the constraint

$$\sum_{i=1}^{n} d_i^2 z_i = \lambda_0 \cdot \sum_i z_i^{p/(p+2)} = c_{\text{total}} \quad \Rightarrow \quad \lambda_0 = c_{\text{total}} \cdot \left(\sum_i z_i^{p/(p+2)}\right)^{-1} \tag{5.32}$$

Plug the optimal d_i^2 and λ_0 back into the original optimization problem, we eventually get the final form of the optimization

$$\underset{\mathbf{U}}{\text{minimize}} \quad \phi^{-1}\left(\sum_{i=1}^{n}\phi(z_i)\right) \tag{5.33}$$

where $\phi(z) = z^{p/(p+2)}$ is a positive and concave function for any $p > 0$. Note that we also have the implicit constraint $\sum_i z_i = \text{tr}\left[\mathbf{U}^T\mathbf{C}\mathbf{U}\right] = \text{tr}\left[\mathbf{C}\right] = const.$ Therefore to minimize the above problem, we need to make the diagonal terms z_i to be different from each other

as much as possible. The extreme case is achieved when $\mathbf{U}^T\mathbf{C}\mathbf{U} = \mathbf{\Lambda}$ is diagonal and z_i's are exactly the eigenvalues up to any permutation. Therefore we can conclude that the optimal $\mathbf{U}^*$ has to diagonalize the input covariance $\mathbf{C}$. For formal proof see Section 5.6.1.

Here we summarize the result for a complete or overcomplete linear population. Let $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ be the eigenvalue decomposition of the stimulus covariance $\mathbf{C}$. Then optimal filter $\mathbf{W}_*$ is

$$\mathbf{W}_* \propto \mathbf{U}_{n\times n}\left[\left(\mathbf{\Lambda}_{n\times n}\right)^{-1/(p+2)}, \mathbf{0}_{n\times(m-n)}\right]\mathbf{V}_{m\times m} \tag{5.34}$$

and the constant scalar is determined by the total energy budget. The linear encoding procedure $\mathbf{t} = \mathbf{W}_*^T\mathbf{s}$ can be summarized as the following. First of all, the input stimulus $\mathbf{s}$ is projected to obtain eigen-components $\mathbf{U}^T\mathbf{s}$. Then depending on the value of p, each eigen-component is renormalized by the matrix $\mathbf{\Lambda}^{-1/(p+2)}$. Next, additional zeros were added to embed this n-dimensional signal into a m-dimensional space to obtain the partially whitened stimulus $\tilde{\mathbf{s}}_p$

$$\tilde{\mathbf{s}}_p = \begin{bmatrix} \mathbf{\Lambda}^{-1/(p+2)}\mathbf{U}^T\mathbf{s} \\ \mathbf{0} \end{bmatrix}. \tag{5.35}$$

At last, a random projection matrix $\mathbf{V}$ is selected to complete the linear filter $\mathbf{t} = \mathbf{V}^T\tilde{\mathbf{s}}_p$. The actual response is subject to some noise $\mathbf{r} = \mathbf{t} + \boldsymbol{\epsilon}$. This process is illustrated in [figure x]. This result provides us some insight on how the optimal encoding strategy varies with the optimal criteria, using a simple linear population codes. For example when $p = 0$, we revisit the infomax solution where the renormalization matrix $\mathbf{\Lambda}^{-1/2}$ is exactly the whitening matrix. Similarly for other values of p, the renormalization will only partially whiten the input stimulus. For example, the L_2 optimal code uses orthogonal filters to process half-whitened data instead of fully whitened the data and the L_∞ optimal code uses orthogonal filters in the original space without any preprocessing. See Figure 18 for an summary of the optimal linear encoding process. In Figure 19 we show the solutions $\mathbf{W}^* = (\mathbf{w}_1, \mathbf{w}_2)$ for

the special case where $m = n = 2$ for $p = 0, 2, 8$.

From the information theoretical perspective, this entire problem for linear neural population can also be understood as a robust coding problem for Gaussian channels. The study of optimal infomax coding has been studied in (Atick. and Redlich, 1990; Guo et al., 2005). A special case for L_2 optimal code in (Doi et al., 2005; Doi and Lewicki, 2011), where the Gaussian prior distribution allows a maximum a posteriori estimator. In this specific case, our L_2 solution is a special case when the signal-noise-ratio (SNR) is large. On the other hand, our solution is valid for all value of p and does not require the prior distribution to be Gaussian.

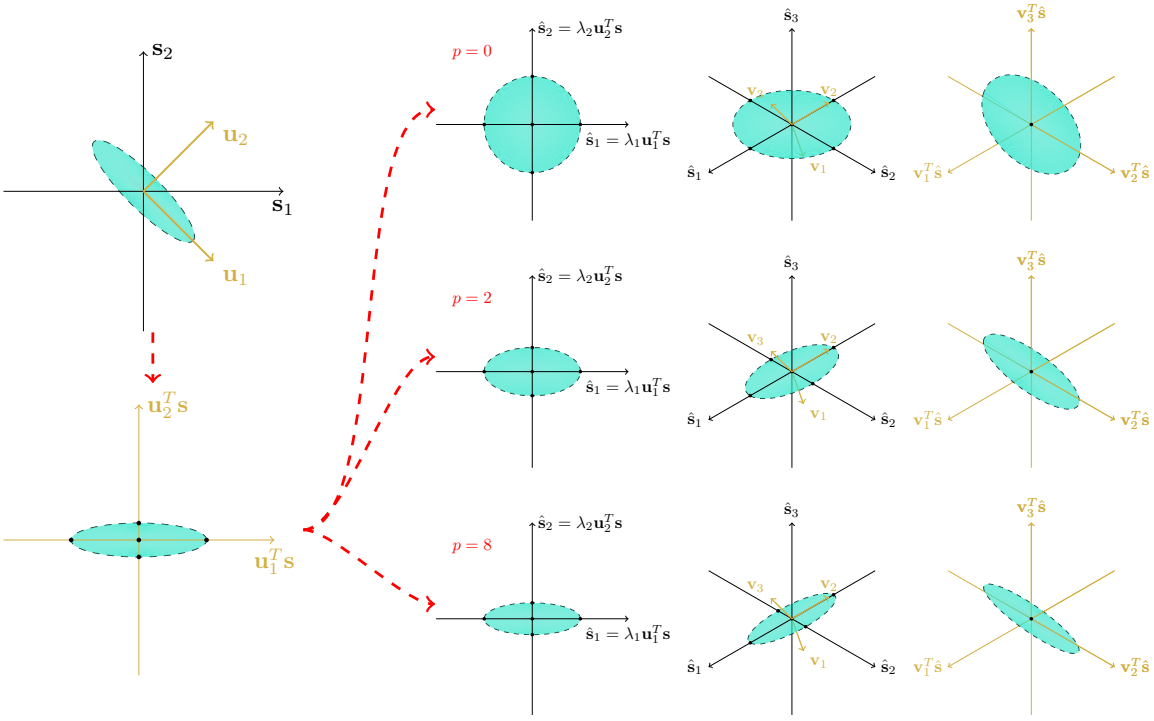

Figure 18: The optimal linear encoder for arbitrary prior distributed stimulus variable. The input stimulus $\mathbf{s}$ is (1) mapped to its eigenspace $\mathbf{U}^T\mathbf{s}$; (2) partially whitened as $\mathbf{\Lambda}^{-1/(2+p)}\mathbf{U}^T\mathbf{s}$ to a degree depending on p; (3) embedded in a higher dimensional space by adding additional zeros $\tilde{\mathbf{s}} = [\mathbf{\Lambda}^{-1/(2+p)}\mathbf{U}^T\mathbf{s}; \mathbf{0}]$; (4) projected by an orthogonal basis to generate the overcomplete representation $\mathbf{t} = \mathbf{V}^T\tilde{\mathbf{s}}$.

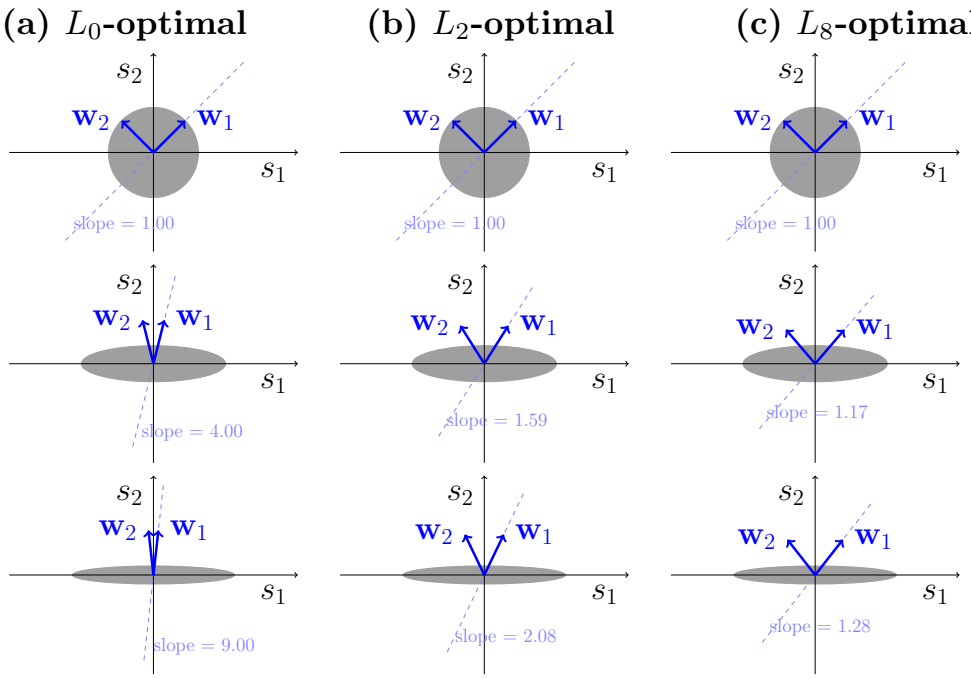

Figure 19: The L_p-optimal filters $\mathbf{W}$ to encode a two-dimensional stimulus variable with certain covariance. In each plot, we show one specific solution $\mathbf{W}^*$ so that $\mathbf{w}_1, \mathbf{w}_2$ are symmetric about y-axis. The aspect ration of the prior distributions betwee eigen-directions varies from 1:1, 4:1 to 9:1. The value of p is 0,2,8 from left to right. Smaller value of p lead to solutions with higher sensitivity to changes in input variance.

5.3. Results for Linear-nonlinear neurons

Now we move on to the general case of using a population of m linear-nonlinear neurons to encode an n-dimensional stimulus variable. The noise model is assumed to be constant Gaussian. As we have discussed, the optimal population with Poisson noise can be easily derived using the optimal population with Gaussian noise. However, we assume the metabolic constraint is imposed on the underlying population of Poisson neurons. In this case, if the gain of each corresponding Gaussian neuron is g_i (for optimization simplicity), then the actual energy constraint should be

$$\mathrm{tr}\left[\mathbf{G}^2\right] \leq g_{\mathrm{total}} \tag{5.36}$$

5.3.1. The Optimization Problem

In order to optimize the left side of Eq. (5.21), the objective function for L_p error minimization is

$$\text{minimize} \qquad L_p(\mathbf{W},\mathbf{P},\mathbf{G},\mathbf{H}) = \left\langle \text{tr}\left[\left(\mathbf{P}^T\mathbf{G}^{-1}\mathbf{H}(\mathbf{s})^{-2}\mathbf{G}^{-1}\mathbf{P}\right)^{p/2}\right]\right\rangle \tag{5.37}$$

$$\text{subject to} \qquad \text{rank}(\mathbf{W}) = n, \quad \mathbf{WP} = \mathbf{I}_n, \quad \text{tr}\left[\mathbf{G}^2\right] \leq g_{\text{total}} \tag{5.38}$$

$$0 \leq h_k(\cdot) \leq 1, \quad k = 1,\ldots,m, \tag{5.39}$$

where $\mathbf{W},\mathbf{P},\mathbf{G},\mathbf{H}$ are defined in Section 5.1. The objective function $L_p(\mathbf{W},\mathbf{P},\mathbf{G},\mathbf{H})$ still cannot be optimized analytically. However, if we artificially limit $\mathbf{P}$ to be not $\mathbf{s}$ dependent, then there are two sharp bounds for the objective function which allow us to obtain a good characterization of near optimal solutions:

$$\text{tr}\left[\left(\mathbf{P}^T\mathbf{G}^{-1}\left\langle\mathbf{H}(\mathbf{s})^{-p}\right\rangle^{2/p}\mathbf{G}^{-1}\mathbf{P}\right)^{p/2}\right] \tag{5.40}$$

$$\leq L_p(\mathbf{W},\mathbf{P},\mathbf{G},\mathbf{H}) \tag{5.41}$$

$$\leq \text{tr}\left[\left(\mathbf{P}^T\mathbf{G}^{-1}\left\langle\mathbf{H}(\mathbf{s})^{-2}\right\rangle\mathbf{G}^{-1}\mathbf{P}\right)^{p/2}\right] \tag{5.42}$$

which is valid for $0 < p \leq 2$. For $p \geq 2$, these inequalities are reversed (see Section 5.6.3 for detailed derivation). It is clear that when $p = 2$, these inequalities hold exactly because the left and right sides are equal. Although we lose the flexibility to choose the optimal $\mathbf{P}$ in Eq. (5.22), but we have derived two bounds which nicely isolate the optimization process of $\mathbf{H}$ from other linear parts. One can analytically minimize either of these two bounds in the general problem described as below

$$\text{minimize} \qquad \text{tr}\left[\left(\mathbf{P}^T\mathbf{G}^{-1}\left\langle\mathbf{H}(\mathbf{s})^{-p}\right\rangle^{2/p}\mathbf{G}^{-1}\mathbf{P}\right)^{p/2}\right] \tag{5.43}$$

$$\text{subject to} \qquad \text{rank}(\mathbf{W}) = n, \quad \mathbf{WP} = \mathbf{I}_n, \quad \text{tr}\left[\mathbf{G}^2\right] \leq g_{\text{total}} \tag{5.44}$$

$$0 \leq h_k(\cdot) \leq 1, \quad k = 1,\ldots,m, \tag{5.45}$$

The above problem can be analytically optimized in two sequential stages. The first stage in Section **??** is to choose the nonlinearity $h_k(\cdot)$ to optimize the matrix $\langle \mathbf{H}^{-p} \rangle$ for any given $\mathbf{W}, \mathbf{P}, \mathbf{G}$. To do this, one can simply minimize each diagonal entry $\left\langle h_k'(\mathbf{w}_k^T \mathbf{s})^{-p} \right\rangle$ because (a) each diagonal element of $\langle \mathbf{H}^{-p} \rangle$ only depends on $h_k'(\cdot)$ and can be minimized individually and (b) for any two positive definite diagonal matrix $\mathbf{D}_1$, $\mathbf{D}_2$ with $(\mathbf{D}_1 - \mathbf{D}_2)_{ii} \geq 0$, the order of the trace power function is preserved:

$$\mathrm{tr}\left[\left(\mathbf{P}^T \mathbf{G}^{-1} \mathbf{D}_1 \mathbf{G}^{-1} \mathbf{P}\right)^{p/2}\right] \geq \mathrm{tr}\left[\left(\mathbf{P}^T \mathbf{G}^{-1} \mathbf{D}_2 \mathbf{G}^{-1} \mathbf{P}\right)^{p/2}\right] \tag{5.46}$$

for any full rank matrix $\mathbf{P}$, positive definite $\mathbf{G}$ and some $p > 0$. The second stage in Section 5.3.3 involves the optimization of the energy budget $\mathbf{G}$ and the linear filter $\mathbf{W}$ by assuming the optimal $\mathbf{H}$ derived from the first stage and certain form of $\mathbf{P}$ to avoid intractability.

5.3.2. Minimizing $\langle \mathbf{H}^{-p} \rangle$

The first step is to explicitly calculate the expectation term which would reduce optimization problem to the one we have dealt with in Section 5.2. Here we minimize each diagonal entry of the matrix $\langle \mathbf{H}^{-p} \rangle$ in Eq. (5.43), assuming some fixed linear filter $\mathbf{W}$ and energy allocation $\mathbf{G}$. We can maximize those diagonal entries one at a time. For each index k, we solve

$$\underset{h_k}{\text{minimize}} \qquad \left\langle h_k'(\mathbf{w}_k^T \mathbf{s})^{-p} \right\rangle = \int f(\mathbf{s}) h_k'(\mathbf{w}_k^T \mathbf{s})^{-p} \, ds \tag{5.47}$$

$$\text{subject to} \qquad 0 \leq h_k(\cdot) \leq 1 \tag{5.48}$$

This problem is equivalent to the one dimensional problem which we have solved in Chapter 4. Here the input of the k-th neuron is the activity generator $t_k = \mathbf{w}_k^T \mathbf{s}$ with some marginal density $f_k(t_k)$. Using the results from Chapter 4, the optimal solution is given by

$$h_k'(t_k) \propto f_k(t_k)^{1/(1+p)} \quad \Rightarrow \quad h_k(t_k) = \frac{\int_{-\infty}^{t_k} f_k(\xi)^{1/(1+p)} \, d\xi}{\int_{-\infty}^{\infty} f_k(\xi)^{1/(1+p)} \, d\xi} \tag{5.49}$$

Once we plug this optimal nonlinearity into the original objective function, it is easy to show that the objective function has the same unit as the variance of its input $t_k = \mathbf{w}_k^T \mathbf{s}$. The minimum value is

$$\left\langle h_k'(\mathbf{w}_k^T \mathbf{s})^{-p} \right\rangle^{2/p} = \left(\int f_k(t_k)^{1/(1+p)} \, dt_k \right)^{(1+p)\cdot 2/p} \tag{5.50}$$

$$= c_p(f_k) \cdot \mathbf{Var}[t_k] = c_p(f_k) \cdot (\mathbf{W}^T \mathbf{C} \mathbf{W})_{kk} \tag{5.51}$$

where the multiplier $c_p(f_k)$ determined by both the marginal distribution f_k and also the optimal criteria p.

Remark 1. For arbitrary prior distribution $f(\mathbf{s})$, it is impossible to derive analytical result because the form of the marginal density $f_k(t_k)$ depends on the filter $\mathbf{w}_k$ in a non-tractable way. However, under certain conditions the problem can still be analytically solved. One such condition is that the prior distribution $f(\mathbf{s})$ is an elliptical distribution. In this case, all one dimensional marginal distribution $f_k(t_k)$ are characterized by a shared template density $f_0(t)$ with unit sample variance and a scaler variable σ_k. In other words,

$$f_k(t_k) = \frac{1}{\sigma_k} \cdot f_0 \left(\frac{t_k}{\sigma_k} \right) \tag{5.52}$$

In this case, the coefficients $c_p(f_k)$ are all equal to $c_p(f_0)$ since all 1D projections have exactly the same marginal distribution once the variance has been normalized out. The matrix $\langle \mathbf{H}^{-p} \rangle^{2/p}$ is proportional to a diagonal matrix with diagonal entries being the variance $(\mathbf{W}^T \mathbf{C} \mathbf{W})_{kk}$ of each 1D projection.

Remark 2. Although there exists an optimal nonlinearity $h_k^*(\cdot)$, it is unclear whether real neurons are capable of achieving this optimality. As in the literature, the emphasis is often put on finding the optimal linear filter $\mathbf{W}$ instead of finding the optimal nonlinearity itself. A generic choice of h_0 is often assumed, such as logistic, error function, fractions of polynomials etc.. If the nonlinearity for each neuron is generated by a fixed nonlinearity h_0

and the rescaling factor σ_k, i.e.

$$h_k(t_k) = \frac{1}{\sigma_k} \cdot h_0\left(\frac{t_k}{\sigma_k}\right) \tag{5.53}$$

With this alternative choice of nonlinearity h_0, it is obvious that Eq. (5.51) will also hold with a suboptimal factor $c_p(f_0, h_0) \geq c_p(f_0, h^*)$.

Remark 3. For different values of p, we have seen that two different optimal solutions of the nonlinearities are derived, each minimize the lower or upper bound of the original objective function, respectively. For general value of p, although it remains unclear where the global optimal solution is, we can still assert that our upper/lower bound minimizer $(\mathbf{W}^*, \mathbf{G}^*, \mathbf{H}^*)$ is near-optimal. And the gap only depends on the parameter p and the one dimensional marginal distribution.

5.3.3. Optimization of $\mathbf{G}$ *and* $\mathbf{W}$

From now on, we assume that the input prior distribution is elliptical (see Remark 1 in Section **??**). By plugging in the optimal value from Eq. (5.51) and dropping the constant, the new optimization problem is

$$\underset{\mathbf{W},\mathbf{P},\mathbf{G}}{\text{minimize}} \qquad \operatorname{tr}\left[\left(\mathbf{P}^T\mathbf{G}^{-1}\mathbf{K}(\mathbf{W})\mathbf{G}^{-1}\mathbf{P}\right)^{p/2}\right] \tag{5.54}$$

$$\text{subject to} \qquad \operatorname{rank}(\mathbf{W}) = n, \quad \mathbf{W}\mathbf{P} = \mathbf{I}_n, \quad \operatorname{tr}\left[\mathbf{G}^2\right] \leq g_{\text{total}} \tag{5.55}$$

where $\mathbf{K}(\mathbf{W})$ is proportional to $\langle \mathbf{H}^{-p} \rangle_*^{2/p}$

$$\mathbf{K}(\mathbf{W}) = \begin{bmatrix} (\mathbf{W}^T\mathbf{C}\mathbf{W})_{11} & & 0 \\ & \ddots & \\ 0 & & (\mathbf{W}^T\mathbf{C}\mathbf{W})_{mm} \end{bmatrix} \tag{5.56}$$

We will show that the remaining problem is almost exactly the same as the problem in Section 5.2. First, we rescale the column of $\mathbf{W}$ by denoting $\mathbf{W} = \tilde{\mathbf{W}}\mathbf{D}$ and we force

$(\tilde{\mathbf{W}}^T\mathbf{C}\tilde{\mathbf{W}})_{kk} = \mathbf{G}_{kk}^2$. In this way, we have $\mathbf{G}^{-1}\mathbf{K}(\mathbf{W})\mathbf{G}^{-1} = \mathbf{D}^2$. In order to make the objective function free of the scaling matrix $\mathbf{D}$, here we pose the additional constraint $\mathbf{P} = \mathbf{D}^{-1}\tilde{\mathbf{W}}(\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T)^{-1}$. With this sub-optimal decoder $\mathbf{P}$, the original problem is equivalent to

$$\underset{\tilde{\mathbf{W}},\mathbf{G}}{\text{minimize}} \qquad \text{tr}\left[\left(\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T\right)^{-p/2}\right] \tag{5.57}$$

$$\text{subject to} \qquad \text{rank}(\tilde{\mathbf{W}}) = n, \quad \text{tr}\left[\mathbf{G}^2\right] \leq g_{\text{total}} \tag{5.58}$$

$$(\tilde{\mathbf{W}}^T\mathbf{C}\tilde{\mathbf{W}})_{ii} = \mathbf{G}_{ii}^2, \quad i = 1, \ldots, n. \tag{5.59}$$

The choice of $\mathbf{G}$ is completely determined by $\tilde{\mathbf{W}}$ and the problem is now exactly what we have solved in Section 5.2!. The total output power is now limited by G_{total}^2. The optimal $\tilde{\mathbf{W}}$ is exactly the same as in Eq. (5.26) and each individual $\mathbf{G}_{ii}$ can be calculated thereafter. With this choice of $\mathbf{P}$, the scalar $\mathbf{D}$ only affects intermediate processing steps in a trivial way but does not affect the neural code quality. For this reason we let $\mathbf{D} = \mathbf{I}$ and replace $\tilde{\mathbf{W}}$ by $\mathbf{W}$.

5.4. Application to Natural Images

Our results can also be applied to higher dimensional stimulus. In this section, we discuss how to build L_p-optimal encoders for natural images. Much work has been done to understand natural images and their impact on the formation of the visual system. For nice and complete review articles, the readers are refer to Simoncelli and Olshausen (2001); Olshausen and Field (2005).

We choose van Hateren's dataset (van Hateren and van der Schaaf, 1998) as the source to generate smaller patches of natural images. Each images in the dataset has 1536x1024 pixels and we shrink its width and height to half (768x512). We apply logarithmic transformation on the raw intensity of each pixel. A total of 50,000 patches of size 8x8 were sampled from random locations of these images and the local mean is removed. Then we stack the 64 pixel values into a 64 dimensional vector which is the high dimensional stimulus to be encoded.

Since all patches are of zero mean, the effective dimension of the stimulus is 63. A few examples of these small patches can be found in Fig.20(a).

5.4.1. Near-Elliptical Prior Distribution

Before we apply any results derived in Section 5.3, we need to confirm that the elliptical assumption of the prior distribution is satisfied for our dataset. The topic of natural images prior distribution has received much attention in the literature and many models have been proposed (Lee et al., 2003; Teh et al., 2003; Sinz and Bethge, 2010; Zoran and Weiss, 2012). Among all the models, the independent component analysis (ICA) model (Comon, 1994) is most closely related to our paper. Traditional ICA model assumes that the high dimensional data is a linear sum of several unknown independent sources. Based on this assumption, structures like localized edges can be recovered as independent sources of natural images (Bell and Sejnowski, 1997). Despite the great similarity between such edge structures and actual neural filters in the primary visual cortex, it has been criticized that the recovered components are not independent (Sinz and Bethge, 2008). To resolve this issue, many efforts have been made to better characterize the prior distribution of natural images. In particular, elliptical distributions seem to be an attractive choice to model the wavelet coefficients for filter pairs close to each other (Wainwright and Simoncelli, 1999; Lyu et al., 2009; Sra et al., 2015).

To confirm the near-elliptical nature of our dataset, we calculate random projections of our image patches by linearly passing these patches through random filters. If a random variable $\mathbf{s}$ follows an elliptical distribution, then marginal distributions of the linear projections $\mathbf{t} = \mathbf{w}^T\mathbf{s}$ can only differ from one aother by a scale parameter. In Fig. 20(b) we illustrate the joint distribution of the output t_1, t_2 of two uncorrelated random filters $\mathbf{w}_1, \mathbf{w}_2$. The conditional distribution $f(t_2|t_1)$ is t_1 dependent, as illustrated in Fig. 20(c).

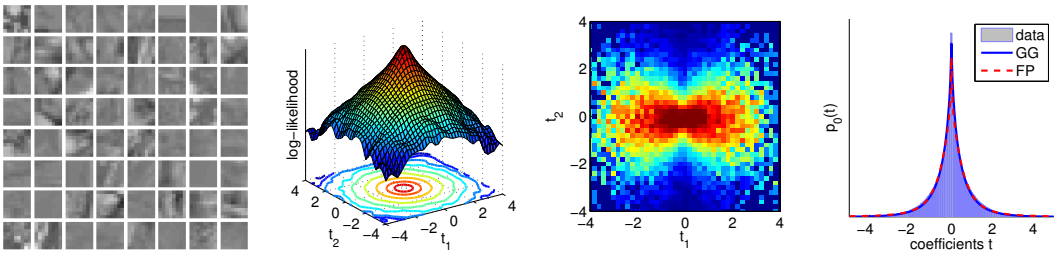

Figure 20: **(a)** 64 out of 50,000 stimulus patches $\mathbf{s}$ randomly sampled from van Hateren's dataset. **(b)** The logarithm of joint 2D-histogram (locally smoothed for clarity) of coefficients t_1 and t_2 where $t_k = \mathbf{w}_k^T \mathbf{s}$ and $\mathbf{w}_k$'s are random filters orthonormal to each other. The joint distribution has spherical contours but is clearly different from 2D Gaussian densities whose log-likelihood are always paraboloidal. **(c)** The conditional distribution $p(t_2|t_1)$ has the "bow-tie" structure. **(d)** Each 1D marginals can be modeled by various parametric models.

5.4.2. Choices of Nonlinearity

For an elliptical distributed stimulus and a fixed (not necessarily the optimal) nonlinearity, we have analytically derived the optimal filters $\mathbf{W}^* = \mathbf{U}^*[\mathbf{\Lambda}^{-1/(2+p)}, \mathbf{0}]\mathbf{V}$ up to an arbitrary unitary matrix $\mathbf{V}$ (see Section 5.3.3). However, the dataset is usually not perfectly symmetric and the one dimensional marginals are slightly different from each other therefore Eq. (5.51) is no longer valid. Since the dataset is still near-elliptical, we still want to assume the new solution does not change much from the original solutions. As the first order perturbation, we assume the optimal solution still takes the form $\mathbf{W}^* = \mathbf{U}^*[\mathbf{\Lambda}^{-1/(2+p)}, \mathbf{0}]\mathbf{V}^*$ but now certain $\mathbf{V}^*$ are superior to other unitary matrices due to the asymmetry. In this section, we study how to find the optimal $\mathbf{V}^*$ based on the dataset.

For the L_0 (infomax) case, the stimulus is first fully whitened and padded with additional zeros to obtain $\tilde{\mathbf{s}}_0$ (see Eq. (5.35)). Then the infomax projection $\mathbf{V}^*$ can be learned by using ICA algorithms (Bell and Sejnowski, 1995; Hyvärinen and Oja, 1997). In this process, it is important to use the correct form of nonlinearity which relies on knowing whether each source has sub-Gaussian or super-Gaussian distributions (Lee et al., 1999a). For the general L_p case, the situation is quite similar. The stimulus variable is first partially-whitened and padded with additional zeros to get $\tilde{\mathbf{s}}_p$ (see Eq. (5.35)). Then we find the best projection

$\mathbf{V}$, which minimizes the total L_p loss on these directions. How such symmetry between $\mathbf{V}$'s breaks down also depends on the form of the assumed nonlinearity.

For any fixed nonlinearity $h_k(\cdot)$, the average L_p loss associated with that single neuron is

$$\langle C_{h_k} \rangle = \int f_k(t_k) C_{h_k}(t_k)\, dt_k \tag{5.60}$$

For brevity, we denote C_{h_k} as the loss function which can be either $-\log h_k'(t)$ when $p = 0$ or $h_k'(t)^{-p}$ when $p > 0$. In Table 1 we illustrate how a single parameter β in the nonlinearities differentiates the sparsity preference for the marginal distributions. In particular, for both criteria when $\beta = 2$, the corresponding nonlinearities are sparsity-neutral. For the infomax ($p = 0$) case, this sparsity-neutral nonlinearity is the error function with derivative proportional to the density of certain Gaussian distribution. Nonlinearities with sub-Gaussian tail (when $\beta < 2$, e.g. logistic function $h'(s) \sim \exp(-|s/\gamma|)$ when s is large) are sparsity-seeking. For the general L_p-min ($p > 0$) case, the sparsity-neutral nonlinearity has derivative $h'(t) \propto (a_0 + a_2|t|^2)^{-1/p}$, which is the density function of a Student-t's distribution.

	Infomax ($p = 0$)		
nonlinearity $h'(t)$	cost function $C_h = -\log h'$	key term in $\langle C_h \rangle$	sparsity preference
$\exp(-a_\beta \lvert t\rvert^\beta)$	$a_\beta \lvert t\rvert^\beta$	$\langle \lvert t\rvert^\beta \rangle$	seeking ($\beta < 2$) neutral ($\beta = 2$) adverse ($\beta > 2$)
	L_p-min ($p > 0$)		
nonlinearity $h'(t)$	cost function $C_h = (h')^{-p}$	key term in $\langle C_h \rangle$	sparsity preference
$\left(a_0 + a_1\lvert t\rvert^2 + a_\beta \lvert t\rvert^\beta\right)^{-1/p}$	$a_0 + a_1\lvert t\rvert^2 + a_\beta \lvert t\rvert^\beta$	$\langle \lvert t\rvert^\beta \rangle$	seeking ($\beta < 2$) neutral ($\beta = 2$) adverse ($\beta > 2$)

Table 1: Examples of nonlinearities where the coefficients $a_0, a_1, a_\beta > 0$. The power $\beta > 0$ determines the preferred filters for each nonlinearity. For the infomax (or the L_p-min) criterion, the derivative of the sparsity-neutral nonlinearity corresponds to a Gaussian density function (or Student-t's density function).

Now we go back to analyze the dataset and need to choose the appropriate nonlinearity to encode the marginal distributions. There are several reasonable choices of parametric

models to describe the marginal distributions. In the infomax scenario, the most widely used model is the Generalized Gaussian (GG) model,

$$\textbf{GG}: \quad f_0(t) \propto \exp(-a_\beta |t|^\beta) \tag{5.61}$$

$$h'(t) \propto f_0(t) \propto \exp(-a_\beta |t|^\beta) \tag{5.62}$$

$$\Rightarrow \quad C_h(t) = -\log h' \propto |t|^\beta \tag{5.63}$$

The nonlinearities in Eq. (5.62) are well understood as sub/super-Gaussian densities depending on the value of β. These nonlinearities can also greatly simplify the computation to calculate the cost. However, such benefit does not extend to other L_p loss function in general. For L_p-min purpose, it is more natural to use the fractional powers of polynomials (FPoP) model below

$$\textbf{FPoP}: \quad f_0(t) \propto (a_0 + a_1|t|^2 + a_\beta |t|^\beta)^{-(1+p)/p} \tag{5.64}$$

$$h'(t) \propto f_0(t)^{1/(1+p)} \propto (a_0 + a_1|t|^2 + a_\beta |t|^\beta)^{-1/p} \tag{5.65}$$

$$\Rightarrow \quad C_h(t) = (h')^{-p} \propto a_0 + a_1|t|^2 + a_\beta |t|^\beta \tag{5.66}$$

In the FPoP model with $\beta < 2$, the associated nonlinearity prefers sparser marginal distributions and the symmetry between unitary matrices $\mathbf{V}$ will break down in a similar way as previous studies on ICA. When applied to the image patches, both GG and FPoP models can achieve comparable data likelihood once the parameters are properly chosen (see Figure 20(d)).

As a remark, we note that the derivative $h'(t)$ in Eq. (5.65) tails off slower than $|t|^{-1}$ when $p \geq 2$. Thus $h'(t)$ does not integrate up to a finite value, which violates the saturation assumption $0 \leq h \leq 1$ for the corresponding nonlinearity $h(t)$. This issue can be partially resolved by setting a cutoff value $t_{\max}$ and let $h'(t) = 0$ for $|t| > t_{\max}$. But for the purpose of evaluating the L_p loss, we will ignore this and just calculate the sample average to approximate the expectation of Eq. (5.66).

5.4.3. Symmetry Breaking for L_p-min Problem

Now we try to find the optimal $\mathbf{V}^*$ to optimize Eq. (5.54) for the proposed nonlinearity $h_0(t)$ and let $h_k(t) = h_0(t/\sigma_k)$ where the scaler $\sigma_k^2 = \mathbf{w}_k^T \mathbf{C} \mathbf{w}_k$ renormalizes each $t_k = \mathbf{w}_k^T \mathbf{s}$ to have unit variance. Now we can calculate that for any $\mathbf{V}$ and the corresponding $\mathbf{W}$, the new expectation in $\langle \mathbf{H}^{-p} \rangle^{2/p}$ slightly deviates from the variance (see Eq. (5.51), Eq. (5.53)) by a term related to $\left\langle |t_k/\sigma_k|^\beta \right\rangle$.

$$\left\langle h_k'(t_k)^{-p} \right\rangle^{2/p} = \sigma_k^2 \cdot \left(a_0 + a_1 \left\langle |t_k/\sigma_k|^2 \right\rangle + a_\beta \left\langle |t_k/\sigma_k|^\beta \right\rangle \right)^{2/p} \tag{5.67}$$

$$\approx \sigma_k^2 \left(a_0 + a_1\right)^{2/p} \cdot \left(1 + (2/p) \frac{a_\beta}{a_0 + a_1} \frac{\left\langle |t_k|^\beta \right\rangle}{\sigma_k^\beta} \right) \tag{5.68}$$

where the expansion is valid when the coefficient $a_\beta \to 0^+$ and the nonlinearity just shifts away from being sparsity-neutral. In order to compensate for this change, the gain of each neuron should also be updated as well $g_k^2 \propto \langle h_k'(t_k)^{-p} \rangle^{2/p}$. Because the total gain is limited $\mathrm{tr}\left[\mathbf{G}^2\right] \le g_{\text{total}}$, it is sufficient to minimize

$$\min \quad \sum_{k=1}^m \left\langle h_k'(t_k)^{-p} \right\rangle^{2/p} = (a_0 + a_1)^{p/2} \left(\sum_{k=1}^m \sigma_k^2 + (2/p) \frac{a_\beta}{a_0 + a_1} \sum_{k=1}^m \sigma_k^{2-\beta} \left\langle |t_k|^\beta \right\rangle \right) \tag{5.69}$$

Becase $\sum \sigma_k^2 = \mathrm{tr}\left[\mathbf{W}^T \mathbf{C} \mathbf{W}\right] = const$, it is equivalent to just minimize

$$\min \quad \sum_{k=1}^m \sigma_k^{2-\beta} \left\langle |t_k|^\beta \right\rangle \quad \Leftrightarrow \quad \min \quad \sum_{k=1}^m (\mathbf{w}_k^T \mathbf{C} \mathbf{w}_k)^{2-\beta} \cdot \left\langle |\mathbf{w}_k^T \mathbf{s}|^\beta \right\rangle \tag{5.70}$$

by finding the proper directions $\mathbf{w}_k$. If we further use the notation in the partially-whitened space: $\tilde{\mathbf{s}}_p = [\mathbf{\Lambda}^{-1/(2+p)} \mathbf{U}^T \mathbf{s}; \mathbf{0}]$ and $\tilde{\mathbf{C}} = \mathbf{cov}[\tilde{\mathbf{s}}_p]$, then an equivalent problem on the unitary matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ is

$$\sum_{k=1}^m (\mathbf{v}_k^T \tilde{\mathbf{C}} \mathbf{v}_k)^{2-\beta} \cdot \left\langle |\mathbf{v}_k^T \tilde{\mathbf{s}}_p|^\beta \right\rangle \tag{5.71}$$

Notice that the symmetry breaking problem for the unitary matrix $\mathbf{V}$ shares the same format for all values of p. The only difference is how "whitened" the processed stimulus $\tilde{\mathbf{s}}_p$ is. In the canonical infomax case, the data is fully whitened so that the whitened stimulus $\tilde{\mathbf{s}}_{p=0}$ has identity covariance $\tilde{\mathbf{C}}$. Therefore the first term $\mathbf{v}_k^T \tilde{\mathbf{C}} \mathbf{v}_k$ is a constant for any unit vector $\mathbf{v}_k$. The symmetry breaking problem matches exactly with its previous description in the ICA literature (Hyvärinen and Oja, 1997) and the expectation $\left\langle |\mathbf{w}_k^T \mathbf{s}| \right\rangle$ is often replaced by a differentiable function, e.g. $\left\langle \log\cosh(\mathbf{w}_k^T \mathbf{s}) \right\rangle$. Similarly, our generalized problem can also be solved efficiently using gradient descent method.

Using the above method we can train the optimal unitary matrix $\mathbf{V}$ for each value of p. The value β is set to be 1 for simplicity. In Figure 21 we compare the optimal over-complete populations for various value of p, where we optimized 100 neurons to encode the 63 dimensional variables of pixel values in the image patches. Assuming sparsity-seeking nonlinearities, the optimal linear components are also edge-like filters, just as the traditional ICA algorithm for the $p = 0$ case. Due to the edge-like nature of these filters, each of these components can be well described by a Gabor function of certain center (x_0, y_0), edge orientation θ, frequency f, phase ϕ and a Gaussian mask described by σ_x and σ_y:

$$g(x, y|\theta, \sigma_x, \sigma_y, f, \phi) = \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right) \cos(2\pi f x' + \phi) \tag{5.72}$$

$$\text{where} \quad x' = (x - x_0)\cos\theta + (y - y_0)\sin\theta, \quad y' = -(x - x_0)\sin\theta + (y - y_0)\cos\theta. \tag{5.73}$$

Next we compare each L_p-optimal population via the statistics of edge orientation θ, edge wavelength $1/f$ and filter area $\sigma_x \sigma_y$. For all values of p, these populations are concentrated on vertical or near vertical edges ($\theta \approx 90^\circ$). We also do not observe a significant difference for the filter size statistics of different population with different value of p. For the wavelength, however, there is a clear shift in the concentration from low frequency edges towards high frequency edges as we increase p. We speculate that this is because that L_p-optimal population with larger p places stronger emphasis on encoding stimulus with smaller variation but infomax population tends to filter these component out with linear

projections.

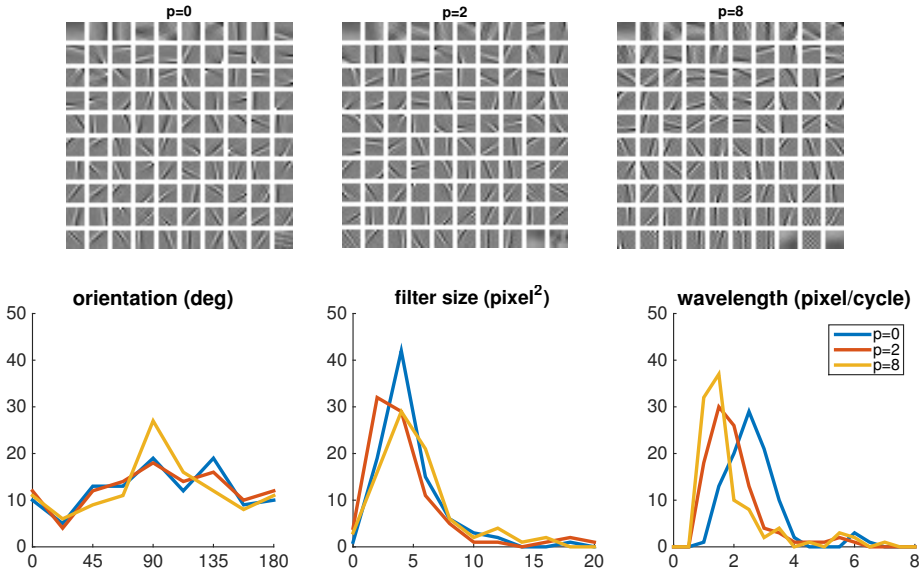

Figure 21: **(a)** 100 Linear components trained for $p = 0, 2, 8$. Each component is fitted by a Gabor function. **(b)** The histograms of orientation parameter θ, filter size $\sigma_x\sigma_y$ and wavelength (inverse frequency) $1/f$ for all neurons in each population.

5.5. Conclusion

Here we summarize the results on optimal population of linear-nonlienar neurons to encode random stimulus which follows an elliptical distribution. The optimal solution consists of successive linear filter part and nonlinear activation part. Under certain limitation, the linear part is given exactly as the linear population case (see Section 5.2) and the nonlinear activation function for each neuron follows the same principle as the one dimensional case (see Chapter 4).

For the complete population case, the optimal linear filter is $\mathbf{W}^* = \mathbf{U}\mathbf{\Lambda}^{-1/(p+2)}\mathbf{V}$. In the above equation, $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is the eigen-decomposition of the data covariance $\mathbf{C} = \mathbf{cov}(\mathbf{s})$ and $\mathbf{V}$ is some arbitrary unitary matrix. For the over-complete population case, the optimal linear filter $\mathbf{W}^* = \mathbf{U}[\mathbf{\Lambda}^{-1/(p+2)}, \mathbf{0}]\mathbf{V}$ is derived by assuming a sub-optimal decoder $\mathbf{P}$. In both cases, the optimal gain for each neuron is $\mathbf{G}^*_{kk} \propto (\mathbf{W}^T\mathbf{C}\mathbf{W})^{1/2}_{kk}$. The optimal nonlinear activation function h_k is determined in the same way as the single neuron case

(see Section 4.2). The only difference is that now the input for each neuron is the activity generator $t_k = \mathbf{w}_k^T \mathbf{s}$ and the L_p-optimal tuning curves are chosen to optimally encode the marginal distribution $f_k(t_k)$.

Again, we remind the reader that the above analytic result is only a near-optimal solution of the L_p-loss minimization problem described in Eq. (5.39), unless $p = 0$ or 2. This situation is summarized in the following table and the gap between the bounds and the actual L_p loss depends on the 1D marginals of the prior distribution (see Remark.3 in Section 5.3.2). Furthermore, for over-complete cases, the decoder is forced to take a sub-optimal choice to obtain this result.

p value	type of optimality	our solution is minimizing
$p = 0$	global optimal	overall L_0-loss (infomax)
$0 < p < 2$	near-optimal	a lower bound of the L_p loss (see Eq. (5.42))
$p = 2$	global optimal	overall L_2-loss (MMSE)
$p > 2$	near-optimal	an upper bound of the L_p loss (see Eq. (5.42))

Table 2: Types of optimality depending on the value of p.

Our results include a unitary symmetry because the prior distribution is assumed to be perfectly elliptical. If we apply the result to datasets with near elliptical prior distribution, we can expect the optimal form of the solution remains the same but the symmetry breaks down. In Section 5.4 we show an application on natural images. Once the symmetry breaks down and the optimal unitary $\mathbf{V}_*$ is found, our result is comparable to many previous results in the literature, in particular the ICA results for complete representation (Bell and Sejnowski, 1995, 1997) or overcomplete representation (Lee et al., 1999b; Lewicki and Sejnowski, 2000). Alternatively, similar results can be obtained by posing sparsity constraint (Olshausen and Field, 1996; Lee et al., 2007) or metabolic cost (Karklin and Simoncelli, 2011).

5.6. Appendix

5.6.1. Proof for Optimization Problem in Eq. (5.33)

Lemma 5.6.1. *Let* $\mathbf{K}$ *be an n-by-n positive definite matrix and* $\phi(z)$ *be a strictly concave function. Then the optimal unitary matrix* $\mathbf{U}$ *should diagonalize* $\mathbf{K}$ *to minimize*

$$\sum_{i=1}^{n} \phi\left((\mathbf{U}^T\mathbf{K}\mathbf{U})_{ii}\right) \tag{5.74}$$

Proof. First we prove this for $n = 2$. For 2-by-2 matrix $\mathbf{K}$, let the two eigenvalues be $0 < \lambda_1 < \lambda_2$. It is obvious that for any unitary matrix $\mathbf{U}$,

$$\lambda_1 \leq (\mathbf{U}^T\mathbf{K}\mathbf{U})_{ii} \leq \lambda_2 \tag{5.75}$$

$$(\mathbf{U}^T\mathbf{K}\mathbf{U})_{11} + (\mathbf{U}^T\mathbf{K}\mathbf{U})_{22} = \lambda_1 + \lambda_2 \tag{5.76}$$

Therefore one can write $(\mathbf{U}^T\mathbf{K}\mathbf{U})_{ii}$ as linear combination of λ_1, λ_2 as

$$(\mathbf{U}^T\mathbf{K}\mathbf{U})_{11} = \alpha\lambda_1 + (1-\alpha)\lambda_2 \tag{5.77}$$

$$(\mathbf{U}^T\mathbf{K}\mathbf{U})_{22} = (1-\alpha)\lambda_1 + \alpha\lambda_2 \tag{5.78}$$

for some $0 \leq \alpha \leq 1$. For concave function $\phi(z)$, it follows from Jensen's inequality that

$$\phi(\alpha\lambda_1 + (1-\alpha)\lambda_2) \geq \alpha\phi(\lambda_1) + (1-\alpha)\lambda_2 \tag{5.79}$$

$$\phi((1-\alpha)\lambda_1 + \alpha\lambda_2) \geq (1-\alpha)\phi(\lambda_1) + \alpha\lambda_2 \tag{5.80}$$

$$\Rightarrow \sum_{i=1,2} \phi\left((\mathbf{U}^T\mathbf{K}\mathbf{U})_{ii}\right) \geq \phi(\lambda_1) + \phi(\lambda_2). \tag{5.81}$$

The equality is attained only if $\alpha = 0, 1$ for strictly concave function ϕ. Therefore to minimize the objective function, $\mathbf{U}$ should be chosen to diagonalize the positive definite matrix $\mathbf{K}$.

In general for the case of $n > 2$, we prove by contradiction. Assume some solution $\mathbf{U}_*$ optimizes the objective function but $(\mathbf{U}_*^T\mathbf{K}\mathbf{U}_*)_{ij} \neq 0$ for some $i \neq j$. Now we consider the two dimensional subspace generated by the i-th and j-th row/column. An additional $\mathbf{U}'$ can be chosen which only diagonalize this 2-by-2 submatrix. For such $\mathbf{U}'$, it does not affect other diagonal entries $(\mathbf{U}_*^T\mathbf{K}\mathbf{U}_*)_{kk}$ for $k \neq i, j$. However it changes $(\mathbf{U}_*^T\mathbf{K}\mathbf{U}_*)_{ii}$ and $(\mathbf{U}_*^T\mathbf{K}\mathbf{U}_*)_{jj}$ but can still improve the objective function (as discussed in the $n = 2$ case), which contradicts the earlier assumption.

□

5.6.2. Preliminary Results on Matrices

Lemma 5.6.2. *Let* $\mathbf{K}$ *be any positive definite matrix. Then for any orthogonal matrix* $\mathbf{U}$ *and any real power* p*, we have*

$$\mathrm{tr}\left[(\mathbf{U}\mathbf{K}\mathbf{U}^T)^p\right] = \mathrm{tr}\left[\mathbf{U}\mathbf{K}^p\mathbf{U}^T\right] = \mathrm{tr}\left[\mathbf{K}^p\right] \tag{5.82}$$

Proof. Obvious as stated in the lemma. □

Lemma 5.6.3. *Let* $\mathbf{A}_{n\times m}$ *be any matrix. Then for any positive power* p*,*

$$\mathrm{tr}\left[(\mathbf{A}\mathbf{A}^T)^p\right] = \mathrm{tr}\left[(\mathbf{A}^T\mathbf{A})^p\right] \tag{5.83}$$

Proof. Consider the singular value decomposition $\mathbf{A}_{n\times m} = \mathbf{U}_{n\times n}\mathbf{D}_{n\times m}\mathbf{V}_{m\times m}$. Plug this in both sides of the equation we get

$$\mathrm{tr}\left[(\mathbf{A}\mathbf{A}^T)^p\right] = \mathrm{tr}\left[(\mathbf{U}\mathbf{D}\mathbf{D}^T\mathbf{U}^T)^p\right] = \mathrm{tr}\left[(\mathbf{D}\mathbf{D}^T)^p\right] \tag{5.84}$$

$$= \sum_{i=1}^{\min\{m,n\}} ((\mathbf{D}_{ii})^2)^p = \mathrm{tr}\left[(\mathbf{D}^T\mathbf{D})^p\right] = \mathrm{tr}\left[(\mathbf{V}^T\mathbf{D}^T\mathbf{D}\mathbf{V})^p\right] = \mathrm{tr}\left[(\mathbf{A}^T\mathbf{A})^p\right] \tag{5.85}$$

□

Theorem 5.6.4 (Araki-Lieb-Thirring Ineqaulity)**.** *Let* $\mathbf{K}$*,* $\mathbf{M}$ *be any two positive semidef-*

inite matrices. Then

$$\mathrm{tr}\left[(\mathbf{KMK})^p\right] \geq \mathrm{tr}\left[\mathbf{K}^p\mathbf{M}^p\mathbf{K}^p\right] \quad \textit{when } 0 < p \leq 1 \tag{5.86}$$

$$\mathrm{tr}\left[(\mathbf{KMK})^p\right] \leq \mathrm{tr}\left[\mathbf{K}^p\mathbf{M}^p\mathbf{K}^p\right] \quad \textit{when } p \geq 1 \tag{5.87}$$

Proof. See (Lieb and Thirring, 1976; Araki, 1990). □

5.6.3. Proof of Bounds in Section 5.3

Here we prove the bounds in Eq. (5.42). We seek reasonable upper and lower bounds for the following quantity

$$\left\langle \mathrm{tr}\left[\left(\mathbf{PG}^{-1}\mathbf{H}(\mathbf{s})^{-2}\mathbf{G}^{-1}\mathbf{P}^T\right)^{p/2}\right]\right\rangle \tag{5.88}$$

where we assume $\mathbf{P}$ is a projection matrix which does not depend on $\mathbf{s}$.

Case I ($0 < q \leq 2$) Lower Bound:

First we derive a lower bound for the objective function. We can show that

$$\left\langle \mathrm{tr}\left[\left(\mathbf{PG}^{-1}\mathbf{H}(\mathbf{s})^{-2}\mathbf{G}^{-1}\mathbf{P}^T\right)^{p/2}\right]\right\rangle \tag{5.89}$$

$$= \left\langle \mathrm{tr}\left[\left(\mathbf{H}(\mathbf{s})^{-1}\mathbf{G}^{-1}\mathbf{P}^T\mathbf{P}\mathbf{G}^{-1}\mathbf{H}(\mathbf{s})^{-1}\right)^{p/2}\right]\right\rangle \tag{5.90}$$

$$= \left\langle \mathrm{tr}\left[\left(\mathbf{H}(\mathbf{s})^{-1}\mathbf{D}^{-1}\cdot\mathbf{D}\mathbf{G}^{-1}\mathbf{P}^T\mathbf{P}\mathbf{G}^{-1}\mathbf{D}\cdot\mathbf{D}^{-1}\mathbf{H}(\mathbf{s})^{-1}\right)^{p/2}\right]\right\rangle \tag{5.91}$$

$$\geq \left\langle \mathrm{tr}\left[\mathbf{H}(\mathbf{s})^{-p}\mathbf{D}^{-p}\cdot(\mathbf{D}\mathbf{G}^{-1}\mathbf{P}^T\mathbf{P}\mathbf{G}^{-1}\mathbf{D})^{p/2}\right]\right\rangle \tag{5.92}$$

$$= \mathrm{tr}\left[\left\langle\mathbf{H}(\mathbf{s})^{-p}\right\rangle\mathbf{D}^{-p}\cdot(\mathbf{D}\mathbf{G}^{-1}\mathbf{P}^T\mathbf{P}\mathbf{G}^{-1}\mathbf{D})^{p/2}\right] \tag{5.93}$$

where $\mathbf{D}$ is an arbitrary positive definite diagonal matrix which does not depend on $\mathbf{s}$. The first equation is due to Lemma 5.6.3 ; the second equation is by inserting $\mathbf{D}^{-1}\mathbf{D}$ which is an identity matrix; the third inequality follows from Theorem 5.6.4 . Because this lower bound works for any $\mathbf{D}$, we can choose $\mathbf{D} = \langle\mathbf{H}(\mathbf{s})^{-p}\rangle^{1/p}$ so that the first term inside the

trace operator is reduced to identity and the lower bounds is

$$\left\langle \mathrm{tr}\left[\left(\mathbf{P}\mathbf{G}^{-1}\mathbf{H}(\mathbf{s})^{-2}\mathbf{G}^{-1}\mathbf{P}^{T}\right)^{p/2}\right]\right\rangle \tag{5.94}$$

$$\geq \mathrm{tr}\left[(\left\langle \mathbf{H}(\mathbf{s})^{-p}\right\rangle^{1/p}\mathbf{G}^{-1}\mathbf{P}^{T}\mathbf{P}\mathbf{G}^{-1}\left\langle \mathbf{H}(\mathbf{s})^{-p}\right\rangle^{1/p})^{p/2}\right] \tag{5.95}$$

$$= \mathrm{tr}\left[\left(\mathbf{P}\mathbf{G}^{-1}\left\langle \mathbf{H}(\mathbf{s})^{-p}\right\rangle^{2/p}\mathbf{G}^{-1}\mathbf{P}^{T}\right)^{p/2}\right] \tag{5.96}$$

where we applied Lemma 5.6.3 again.

Case I ($0 < q \leq 2$) Upper Bound:

On the other hand, consider the concave operator [cite: theorem 2.10 Eric Carlen] on positive definite matrices $f(\mathbf{M}) = \mathrm{tr}\left[\mathbf{M}^{p/2}\right]$ for $0 < p \leq 2$. Apply Jensen's inequality $\langle f(\mathbf{M})\rangle \leq f(\langle \mathbf{M}\rangle)$ and plug in $\mathbf{M} = \mathbf{P}\mathbf{G}^{-1}\mathbf{H}(\mathbf{s})^{-2}\mathbf{G}^{-1}\mathbf{P}^{T}$, we have

$$\left\langle \mathrm{tr}\left[\left(\mathbf{P}\mathbf{G}^{-1}\mathbf{H}(\mathbf{s})^{-2}\mathbf{G}^{-1}\mathbf{P}^{T}\right)^{p/2}\right]\right\rangle \leq \mathrm{tr}\left[\left(\mathbf{P}\mathbf{G}^{-1}\left\langle \mathbf{H}(\mathbf{s})^{-2}\right\rangle \mathbf{G}^{-1}\mathbf{P}^{T}\right)^{p/2}\right] \tag{5.97}$$

Case II ($q > 2$):

Same as case I, except that all inequalities that have been used are reversed.

5.6.4. The Optimal Non-lienarity and Renyi Entropy

In Section 4.2, we provided the optimal solution $h_*(s)$ for the nonlinearity of a single neuron for given p value, to encode a stimulus with prior density $f(s)$. If we plug this value into the objective function, we can calculate the optimal value

$$\left\langle h_*'(s)^{-p}\right\rangle = \left(\int f(s)^{1/(1+p)}\,ds\right)^{1+p} \tag{5.98}$$

This value is related to the nature of the density function $f(s)$. If we consider the Renyi-α entropy $H_\alpha(s)$ of a distribution $f(s)$ (see Renyi (1961)) and let $\alpha = 1/(1+p)$, then we

immediately have

$$H_\alpha(f) = \frac{1}{1-\alpha} \log \left(\int f(s)^\alpha \, ds \right) = \frac{1}{p} \log \left\langle h'_*(s)^{-p} \right\rangle \tag{5.99}$$

In particular, when we calculate certain power of the optimal value, we have

$$\left\langle h'_*(s)^{-p} \right\rangle^{2/p} = \exp\left(2H_\alpha(f)\right) = c_p(f, h_*) \cdot \mathbf{cov}[s] \tag{5.100}$$

Such value is known as the exponential entropy which has been used to characterize the extent of a distribution (Campbell, 1966). If the distribution f's are from the same family parametrized by a single scale, then the above value is simply proportional to the variance of the distribution.

In equation Eq. (5.99), if we let $p \to 0$ and $\alpha \to 1$, then the optimal solution $h_*(s)$ converges to the (Shannon) infomax rule $h'_*(s) = f(s)$. Correspondingly, it is well known that the Renyi-1 entropy is exactly the canonical Shannon entropy (Renyi, 1961)

$$H_1(f) = \lim_{\alpha \to 1} H_\alpha(f) = -\int f(s) \log f(s) \, ds \tag{5.101}$$

It has also been proved that the Renyi-α entropy is decreasing as a function of α therefore the Renyi-α entropy may diverge for small enough α (large enough p), especially for prior distribution $f(s)$ with polynomially decaying tails. In this case, all possible nonlinearities $h'(s)$ will produce infinite L_p loss and the optimization problem cannot be solved.

CHAPTER 6 : Conclusion

Understanding how neural codes adapt to the sensory stimulus statistics has been a fundamental goal in sensory neuroscience. As one of the best known and most successful theories, the efficient coding hypothesis assumes that biological sensory systems should maximize the information being transferred from the stimulus to the neural output. In this thesis, we have followed, investigated and extended this approach in multiple but systematic ways – we have analytically derived the optimal codes for stimuli with an arbitrary prior, in most cases.

When formulating an efficient coding problem, two key components must be considered: the constraints and the objective function. Previous works often choose a specific way to set up and solve the problem. Here we have presented a unified framework and show that these results can be understood from a much more generalized perspective. Compared to a canonical work by Laughlin (1981), we consider multiple directions to extend the current model. In Chapter 3, we consider the inclusion of a new biologically plausible constraint, which limits the mean activity of neural output. In Chapter 4, we use the traditional range constraint but the L_p metric instead of the mutual information criterion to measure the quality of neural codes. In Chapter 5, we further extend the idea of L_p optimal code to multivariate input. This is summarized in Table 3 and more detailed discussion on each chapter can be found below.

	constraint	objective function	dimension of stimulus $\mathbf{s}$ (n) number of neurons (m)
Laughlin (1981)	range	infomax	$n = 1$ $m = 1$
Chapter 3	range metabolic	infomax	$n = 1$ $m \geq 1$
Chapter 4	range	infomax L_p-optimal	$n = 1$ $m \geq 1$
Chapter 5	range	infomax L_p-optimal	$n > 1$ $m \geq n$

Table 3: Summary of our contribution in this thesis.

In Chapter 3, we presented a theoretical framework for studying optimal neural codes under biologically relevant constraints. Especially, we emphasized the importance of two constraints – the noise characteristics of the neural responses and the metabolic cost. We demonstrated that, maybe surprisingly, analytical solutions exist for a wide family of noise characteristics and metabolic cost functions. This result helps us to determine the optimal tuning curves for multiple neurons and suggests that ON-OFF pathway splitting is superior than ON-ON code only if the metabolic constraint is included in the picture. In our analysis, we have ignored several important other factors when formulating the efficient coding problem. First, we have not modeled the spontaneous activity (baseline firing rate) of neurons. Second, we have only considered the zero noise correlations between the responses of neurons. Third, we have ignored the noise in the input to the neurons. Including these factors should allow us to make a more detailed and quantitative comparison to physiologically measured data in the future.

In Chapter 4, we switched to a framework that generalizes both the mutual information criterion and the square decoding error criterion. We systematically evaluate different optimality criteria based upon the L_p reconstruction error of the maximum likelihood decoder. This parametric family of optimal criteria includes two aforementioned special cases – $p \to 0$ corresponds to the information criterion and $p = 2$ corresponds to the square decoding error criterion. We analytically derived the optimal codes that minimizes the L_p reconstruction error of an ideal observer. Our framework offers greater flexibility when used to explain physiology data. After all, maximizing mutual information may not (and should not) be the only goal of neural codes. Assuming different combination of objective function and constraints, we tested our analytical predictions against previously measured characteristics of some early visual systems found in biology. We find low values of p provides a better fit for physiology data on early visual perception systems.

In Chapter 5, we further extended our previous results by generalizing to multivariate input stimulus. We also considered the neural codes that minimize the L_p reconstruction error

of the stimulus and derived analytical solutions under a few extra assumptions. Similar to before, this framework unifies the formerly well known information maximization criterion ($p \to 0$) and the square decoding loss criterion ($p = 2$). Within our framework, we obtained L_p optimal neural codes for natural image stimuli and found similar edge-like filters as reported in previous results which took the information criterion. Compared to easier setups, the results on multivariate input stimulus has several limitations. To optimize a linear population, we require the prior distribution has finite variance. For a linear-nonlinear population, analytical solutions exist when the prior distribution is elliptical or near-elliptical. Furthermore, if the population is overcomplete $m > n$ or $p \neq 0, 2$, only the near-optimal solution can be obtained because the objective function has to be approximately evaluated.

There are also generic limitations for all chapters. First we did not consider a grand unification of different aspects – the combination of L_p optimal criteria and metabolic constraints. This seems intractable but is an interesting open question for future study. Second, we have derived all results under the low noise assumption which may not be the case in the real world. This limitation can partially be compensated by having sufficient encoding time and/or sufficient replicated neurons performing the same task with independent noise. We also investigated what would happen if some of these key assumptions were removed in Section 4.4. Third, the L_p optimal nonlinearity may diverge for those priors (or one dimensional marginals in multivariate case) with heavy tails. This cannot be resolved because any tuning curve with finite range will all have infinite L_p loss.

Despite these limitations, our analysis is useful in many ways. We simultaneously considered the biologically plausible range constraint and the metabolic constraint, and derive an analytical optimal solution. Most importantly, we proposed a framework which smoothly interpolates the information criterion and the square decoding loss criterion to provide a full family of optimal criteria that can possibly be employed by neural populations in actual perceptual system. Our result provides predictions that can potentially be examined by physiology experiments.

BIBLIOGRAPHY

H. Araki. On an inequality of lieb and thirring. *Letters in Mathematical Physics*, 19(2): 167–170, 1990. ISSN 0377-9017. doi: 10.1007/BF01045887.

J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251, 1992.

J. Atick. and A. Redlich. Towards a Theory of Early Visual Processing. *Neural Computation*, 2(3):308–320, 1990. ISSN 0899-7667. doi: 10.1162/neco.1990.2.3.308.

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

D. Attwell and S. Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of cerebral blood flow and metabolism*, 21(10):1133–1145, 2001. ISSN 0271-678X. doi: 10.1097/00004647-200110000-00001.

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

H. Barlow. *Possible principles underlying the transformation of sensory messages.* M.I.T. Press, 1961.

H. Barlow. Redundancy reduction revisited. *Network: computation in neural systems*, 12 (3):241–253, 2001.

A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

A. Bell and T. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

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

M. Bethge, D. Rotermund, and K. Pawelzik. Optimal short-term population coding: when Fisher information fails. *Neural Computation*, 14:2317–2351, 2002.

M. Bethge, D. Rotermund, and K. Pawelzik. Optimal neural rate coding leads to bimodal firing rate distributions. *Netw. Comput. Neural Syst.*, 14:303–319, 2003.

N. Brady and D. Field. Local contrast in natural images: Normalisation and coding efficiency. *Perception*, 29(9):1041–1055, 2000. ISSN 03010066. doi: 10.1068/p2996.

N. Brenner, W. Bialek, and R. de Ruyter van Steveninck. Adaptive rescaling maximizes information transmission. *Neuron*, 26:695–702, 2000.

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

L. Campbell. Exponential entropy as a measure of extent of a distribution. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 5(3):217–225, 1966. ISSN 0044-3719. doi: 10.1007/BF00533058.

M. Carandini and D. Heeger. Normalization as a canonical neural computation. *Nature Review Neuroscience*, 13:51–62, 2012. ISSN 1471-003X. doi: 10.1038/nrn3136.

E. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network (Bristol, England)*, 12(2):199–213, 2001. ISSN 0954-898X.

M. Churchland et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3):369–378, 2010.

P. Clatworthy, M. Chirimuuta, J. Lauritzen, and D. Tolhurst. Coding of the contrasts in natural images by populations of neurons in primary visual cortex (V1). *Vision Research*, 43(18):1983–2001, 2003. ISSN 00426989. doi: 10.1016/S0042-6989(03)00277-3.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

R. de Ruyter van Steveninck, G. Lewen, and W. Bialek. Reproducibility and variability in neural spike trains. *Science*, 275:1805, 1997.

I. Dean, N. Harper, and D. McAlpine. Neural population coding of sound level adapts to stimulus statistics. *Nature neuroscience*, 8:1684–1689, 2005.

E. Doi and M. Lewicki. Characterization of minimum error linear coding with sensory and neural noise. *Neural Computation*, 23(10):2498–2510, 2011.

E. Doi, D. Balcan, and M. Lewicki. A theoretical analysis of robust coding over noisy overcomplete channels. *Advances in Neural Information Processing Systems*, 18:307–314, 2005.

D. Dong and J. Atick. Statistics of Natural Time-Varying Images. *Network: Computation in Neural System*, 6(3):345–358, 1995.

D. Fitzpatrick, R. Batra, T. Stanford, and S. Kuwada. A neuronal population code for sound localization. *Nature*, 388:871–874, 1997.

D. Ganguli and E. Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. *Adv. Neural Information Processing Systems*, 23:658–666, 2010.

D. Ganguli and E. Simoncelli. Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, 26(10):2103–2134, 2014.

A. Girshick, M. Landy, and E. Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, 14(7):926–932, 2011. ISSN 1097-6256. doi: 10.1038/nn.2831.

J. Gjorgjieva, H. Sompolinsky., and M. Meister. Benefits of Pathway Splitting in Sensory Coding. *Journal of Neuroscience*, 34(36):12127–12144, 2014. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1032-14.2014.

R. Goris, J. Movshon, and E. Simoncelli. Partitioning neuronal variability. *Nature neuroscience*, 17(6):858–65, 2014. ISSN 1546-1726. doi: 10.1038/nn.3711.

A. Gottschalk. Derivation of the visual contrast response function by maximizing information rate. *Neural computation*, 14(3):527–542, 2002.

A. Grabska-Barwinska and J. Pillow. Optimal prior-dependent neural population codes under shared input noise. *Adv. Neural Information Processing Systems*, 27:1880–1888, 2014.

D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. ISSN 00189448. doi: 10.1109/TIT.2005.844072.

N. Harper and D. McAlpine. Optimal neural population coding of an auditory spatial cue. *Nature*, 430:682–686, 2004.

G. Henry, B. Dreher, and P. Bishop. Orientation of Cells in Cat Striate. *Journal of neurophysiology*, 37(6):1394–1409, 1974.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis of complex-valued signals. *Neural Computation*, 9:1483–1492, 1997. doi: 10.1142/S0129065700000028.

D. Johnson and W. Ray. Optimal stimulus coding by neural populations using rate code. *J. Comput. Neurosci*, 16:129–138, 2004.

K. Kang, R. Shapley, and H. Sompolinsky. Information tuning of populations of neurons in primary visual cortex. *Journal of neuroscience*, 24(15):3726–3735, 2004.

Y. Karklin and E. Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. *Advances in Neural Information Processing Systems (NIPS*11)*, 24, 2011.

D. Kastner, S. Baccus, and T. Sharpee. Critical and maximally informative encoding between neural populations in the retina. *Proceedings of the National Academy of Sciences*, 112(8):2533–2538, 2015.

D. C. Knill and W. Richards. *Perception As Bayesian Inference*. Cambridge University Press, New York, NY, USA, 1996. ISBN 0-521-46109-X.

S. Laughlin. A simple coding procedure enhances a neurons information capacity. *Z. Naturforschung*, 36c(3):910–912, 1981.

S. Laughlin, R. Steveninck, and J. Anderson. The metabolic cost of neural information. *Nature neuroscience*, 1(1):36–41, 1998. ISSN 1097-6256. doi: 10.1038/236.

A. B. Lee, K. S. Pedersen, and D. Mumford. The Nonlinear Statistics of High-Contrast Patches in Natural Images. *International Journal of Computer Vision*, 54(1/2/3):83–103, 2003. ISSN 13119109. doi: 10.1023/A.

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. *Advances in Neural Information Processing Systems*, 19:801, 2007. ISSN 10495258. doi: 10.1.1.69.2112.

T.-W. Lee, M. Girolami, and T. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11(2):417–441, 1999a. ISSN 0899-7667. doi: 10.1162/089976699300016719.

T.-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski. Blind source separation of more sources than mixtures using sparse mixture models. *IEEE Signal Processing Letters*, 6 (4):87–90, 1999b. ISSN 0167-8655.

W. Levy and R. Baxter. Energy efficient neural codes. *Neural computation*, 8(3):531–543, 1996.

M. Lewicki and T. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000. ISSN 0899-7667. doi: 10.1162/089976600300015826.

E. Lieb and W. Thirring. Inequalities for the moments of the eigenvalues of the schrodinger hamiltonian and their relation to sobolev inequalities. In W. Thirring, editor, *The Stability of Matter: From Atoms to Stars*, pages 205–239. Springer Berlin Heidelberg, 1976. ISBN 978-3-540-22212-5. doi: 10.1007/3-540-27056-6_16.

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

R. Linsker. An application of the principle of maximum information preservation to linear systems. *Advances Neural Information Processing Systems*, pages 105–117, 1989.

S. Lyu, E. Simoncelli, and H. Hughes. Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Computation*, pages 1485–1519, 2009.

T. Maddess and S. Laughlin. Adaptation of the motion-sensitive neuron h1 is generated locally and governed by contrast frequency. *Proc. R. Soc. Lond. B Biol. Sci*, 225:251–275, 1985.

V. Mante, R. Frazor, V. Bonin, W. Geisler, and M. Carandini. Independence of luminance and contrast in natural scenes and in the early visual system. *Nature neuroscience*, 8(12): 1690–1697, 2005. ISSN 1097-6256. doi: 10.1038/nn1556.

M. McDonnell and N. Stocks. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Phys. Rev. Lett.*, 101:058103, 2008.

J.-P. Nadal and N. Parga. Non linear neurons in the low noise limit: A factorial code maximizes information transfer. *Network: Computation in Neural Systems*, pages 565–581, 1994.

J. Najemnik and W. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 03 2005.

K. Naka and W. Rushton. S-potentials from luminosity units in the retina of fish (cyprinidae). *The Journal of Physiology*, 185(3):587–599, 08 1966.

A. Nikitin, N. Stocks, R. Morse, and M. McDonnell. Neural population coding is optimized by discrete tuning curves. *Phys. Rev. Lett.*, 103:138101, 2009.

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

B. Olshausen and D. Field. How close are we to understanding v1? *Neural computation*, 17(8):1665–1699, 2005. ISSN 0899-7667. doi: 10.1162/0899766054026639.

Y. Ozuysal and S. Baccus. Linking the computational structure of variance adaptation to biophysical mechanisms. *Neuron*, 73:1002–1015, 2012.

S. Pilarski and O. Pokora. On the Cramér-Rao bound applicability and the role of Fisher information in computational neuroscience. *BioSystems*, 136:11–22, 2015. ISSN 18728324. doi: 10.1016/j.biosystems.2015.07.009.

A. Pouget, S. Deneve, J.-C. Ducom, and P. Latham. Narrow versus wide tuning curves: Whats best for a population code? *Neural Computation*, 11:85–90, 1999.

A. Renyi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

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

D. Ringach, R. Shapley, and M. Hawken. Orientation selectivity in macaque V1: diversity and laminar dependence. *Journal of neuroscience*, 22(13):5639–5651, 2002. ISSN 1529-2401. doi: 20026567.

E. Salinas. How behavioral constraints may determine optimal sensory representations. *PLoS Biology*, 4(12):2383–2392, 2006. ISSN 15457885. doi: 10.1371/journal.pbio.0040387.

P. Schiller. The on and off channels of the visual system. *Trends in neurosciences*, 15(3): 86–92, 1992.

H. Seung and H. Sompolinsky. Simple models for reading neuronal population codes. *Proc. of the National Aca. of Sci. of the U.S.A.*, 90:10749–10753, 1993.

C. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

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

E. Simoncelli and B. Olshausen. Natural Image Statistics and Neural Representation. *Annu. Rev. Neurosci.*, 24:1193–1216, 2001.

F. Sinz and M. Bethge. The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In *Advances in Neural Information Processing Systems 21*, pages 1521–1528, 2008.

F. Sinz and M. Bethge. l_p-nested symmetric distributions. *Journal of Machine Learning Research*, 11:3409–3451, Dec 2010.

S. Sra, R. Hosseini, L. Theis, and M. Bethge. Data modeling with the elliptical gamma distribution. In *Artificial Intelligence and Statistics*, volume 18, 2015.

G. Stecker, I. Harrington, and J. Middlebrooks. Location coding by opponent neural populations in the auditory cortex. *PLoS Biology*, 3(3):e78, 2005.

Y. Tadmor and D. Tolhurst. Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision research*, 40(22):3145–3157, 2000. ISSN 00426989. doi: 10.1016/S0042-6989(00)00166-8.

Y. Teh, M. Welling, S. Osindero, and G. Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning*, 4:1235–1260, 2003. ISSN 1532-4435. doi: 10.1162/jmlr.2003.4.7-8.1235.

F. Theunissen and J. Miller. Representation of sensory information in the cricket cercal sensory system. II. information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *J Neurophysiol*, 66:1690–1703, Nov. 1991. ISSN 0022-3077.

G. Tkacik, J. Prentice, V. Balasubramanian, and E. Schneidman. Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 107(32):14419–14424, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1004906107.

D. Tolhurst, J. Movshon, and I. Thompson. The dependence of response amplitude and variance of cat visual cortical neurones on stimulus contrast. *Experimental brain research*, 41(3-4):414–419, 1981.

G. Tomko and D. Crapper. Neuronal variability: non-stationary responses to identical visual stimuli. *Brain research*, 79(3):405–418, 1974.

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

R. D. Valois, E. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22:531–544, 1982.

J. Van Hateren. Spatiotemporal contrast sensitivity of early vision. *Vision Research*, 33(2): 257–267, 1993. ISSN 00426989. doi: 10.1016/0042-6989(93)90163-Q.

J. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings: Biological Sciences*, 265(1394):359–366, Mar 1998.

M. Wainwright and E. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Adv. Neural Information Processing Systems (NIPS*99)*, volume 12, pages 855–861, Cambridge, MA, May 1999. MIT Press.

Z. Wang, A. Stocker, and D. Lee. Optimal neural tuning curves for arbitrary stimulus distributions: Discrimax, infomax and minimum lp loss. *Advances Neural Information Processing Systems*, 25:2177–2185, 2012.

Z. Wang, A. Stocker, and D. Lee. Optimal neural population codes for high-dimensional stimulus variables. *Advances in Neural Information Processing Systems*, 26:297–305, 2013.

H. Wässle. Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5 (10):747–757, 2004.

X.-X. Wei and A. Stocker. A bayesian observer model constrained by efficient coding can explain 'anti-bayesian' percepts. *Nat Neurosci*, 18(10):1509–1517, Oct. 2015. doi: 10.1038/nn.4105.

X.-X. Wei and A. Stocker. Mutual Information, Fisher Information, and Efficient Coding. *Neural Computation*, 28:305–326, 2016.

S. Yaeli and R. Meir. Error-based analysis of optimal tuning functions explains phenomena observed in sensory neurons. *Front Comput Neurosci*, 4, 2010.

S. Yarrow, E. Challis, and P. Seris. Fisher and shannon information in finite neural populations. *Neural Computation*, 24:1740–1780, 2012.

K. Zhang and T. Sejnowski. Neuronal tuning: To sharpen or broaden? *Neural Computation*, 11:75–84, 1999.

L. Zhao and L. Zhaoping. Understanding auditory spectro-temporal receptive fields and their changes with input statistics by efficient coding principles. *PLoS Comput Biol*, 7 (8):1–16, 08 2011. doi: 10.1371/journal.pcbi.1002123.

D. Zoran and Y. Weiss. Natural images, Gaussian mixtures and dead leaves. *Advances Neural Information Processing Systems*, pages 1–9, 2012. ISSN 10495258.