UNIVERSITY OF CALIFORNIA SAN DIEGO

Characterizing and quantifying information conveyed by large neuronal populations

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Physics

by

John Berkowitz

Committee in charge:

Professor Tatyana Sharpee, Chair
Professor Massimiliano Di Ventra, Co-Chair
Professor Terrence Sejnowski
Professor Charles Stevens
Professor Massimo Vergassola

2019

The dissertation of John Berkowitz is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2019

DEDICATION

To my parents, David and Katherine, for supporting and encouraging me along every step of my academic career.

EPIGRAPH

It is not incumbent upon you to complete the work,

but neither are you at liberty to desist from it.

—Ethics of the Fathers (2:21)

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Tatyana Sharpee for her mentorship and guidance. Dr. Sharpee showed me the importance of asking the big questions and not getting hung up on the details. Working in the CNL-T has forced me to go out and "get my hands dirty" as much as any theoretical or computational researcher can.

I would also like to acknowledge the fellow graduate students of the CNL-T, past and present, who have been wonderful and inspiring colleagues. I would particularly like to thank Ryan Rowekamp and Joel Kaardal for their ample help in interacting with our lab's code base. Additionally, CNL technical staff member Jorge Aldana has been of great assistance throughout my time at Salk.

Finally, I would like to thank the members of my committee, Drs. Massimiliano Di Ventra, Terrence Sejnowski, Charles Stevens, and Massimo Vergassola.

Portions of Chapter 2 appear in a manuscript under review. John A. Berkowitz, Tatyana O. Sharpee. Cortical column as an information-preserving decoder of neural inputs. Under review, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reformatted version of the material as it appears in the following publication: John A. Berkowitz, Tatyana O. Sharpee. Quantifying information conveyed by large neuronal populations. Neural Computation, 2019. The dissertation author was the primary investigator and author of this paper.

VITA

2011	B. S. in Physics and Mathematics *cum laude*, Coe College, Cedar Rapids
2011-2018	Graduate Research Assistant, University of California San Diego
2019	Ph. D. in Physics, University of California San Diego

PUBLICATIONS

Lipton, Z. C., **Berkowitz, J.**, Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Berkowitz, J., Sharpee T.O. (2018). Information theoretic constraints on cortical computation. *Under Review*

Berkowitz, J., Sharpee T.O. (2019). Quantifying information conveyed by large neuronal populations. *Neural Computation*

Berkowitz, J. (2019). A Differentiable Approximation of Mutual Information in a Binary Network *In Preparation*

ABSTRACT OF THE DISSERTATION

Characterizing and quantifying information conveyed by large neuronal populations

by

John Berkowitz

Doctor of Philosophy in Physics

University of California San Diego, 2019

Professor Tatyana Sharpee, Chair

Professor Massimiliano Di Ventra, Co-Chair

How neurons in the brain collectively represent stimuli is a long standing open problem. Studies in many species from leech and cricket to primate show that animals behavior,

such as gaze direction or arm movement, often correlates with a measure of neural activity termed the population vector. To construct it, one averages preferred stimuli for each neuron proportionately to their responses. However, the population vector discards much of the information contained in the activity of the population. In the first part of this thesis we show that for a broad class of common models of neural populations a sufficient statistic of the population response can be constructed that is guaranteed to transmit as much information about the stimulus as the population response. The statistic has fixed dimension, independent of the population size, and is valid even in the presence of intrinsic interneuronal correlations. This statistic turns out to be a re-weighted version of the population vector. We validate the performance of this statistic on a dataset of visual neural responses. Additionally we show that under certain conditions, this statistic can serve as a reconstruction of the stimulus itself.

Quantifying mutual information between inputs and outputs of a large neural circuit is an important open problem in both machine learning and neuroscience. However, evaluation of the mutual information is known to be generally intractable for large systems due to the exponential growth in the number of terms that need to be evaluated. Here we show how information contained in the responses of large neural populations can be effectively computed for a class of models that generalize those considered in the first part of the thesis. Neural responses in this model can remain sensitive to multiple stimulus components. We show that the mutual information in this model can be effectively approximated as a sum of lower-dimensional conditional mutual information terms. The approximations become exact in the limit of large neural populations and for certain conditions on the distribution of receptive fields across the neural population. We empirically find that these approximations continue

to work well even when the conditions on the receptive field distributions are not fulfilled. The computing cost for the proposed methods grows linearly in the dimension of the input, and compares favourably with other approximations.

Chapter 1: Introduction

1.1 The Neural Code

A central problem in sensory neuroscience is determining how semantic and task-relevant information about external stimuli is communicated by the outputs of neural populations in lower level sensory areas to higher level decision making parts of an organism's brain. Making progress on understanding this neural code requires answering several related questions such as how to quantify the stimulus, how to characterize the neural outputs, and how to model the variability in neural responses. A common and useful approximation that we assume for the duration of this thesis is that time has been discretized into discrete bins where the stimulus in bin t is $(\vec{s}_t)$ is considered constant and the neural response $(\vec{r}_t)$ recorded over the corresponding bin is considered to be drawn from a stationary stimulus conditioned distribution:

$$\vec{r}_t \sim P_t(\vec{r}) = P(\vec{r}|\vec{s}_t, t) = P(\vec{r}|\vec{s}_t) \tag{1.1}$$

The important approximation implied by (1.1) is that the distribution over $\vec{r}$ depends only on time t through $\vec{s}_t$. This precludes, among other processes, adaptation in spike rate

[Fairhall et al., 2001], rescaling of neural dynamic range for contrast adaptation [Wainwright, 1999, Brenner et al., 2000a, Kastner and Baccus, 2011], as well as certain forms of predictive coding [Hosoya et al., 2005]. While this may seem like a strong simplification, the assumption of stationarity vastly simplifies the analysis of stimulus-response relationships. Additionally, unless otherwise noted we will assume that the stimuli are also drawn from a stationary distribution: $\vec{s}_t \sim P(\vec{s})$. We also have to formalize our representations of the stimulus and the neural responses. The stimulus $\vec{s}$ is modelled simply as a D-dimensional vector in euclidean space. This is a very general representation, where components of $\vec{s}$ could represent the intensity of light coming from pixels on a screen or velocity components in some coordinate frame. Modelling $\vec{r}$ is a more open ended task. We first assume that the population is of a fixed size, denoted N. A very general representation for $\vec{r}$ would be a list of spikes during the time bin, including the time of spiking and index of the neuron that spiked. Including all of this information would be necessary for analysing codes based upon precise spike-timing [Srivastava et al., 2017, Theunissen and Miller, 1995]. For this thesis we instead treat $\vec{r}$ as a length N vector, where r_i represents the number of times the i^{th} neuron spiked during the time bin. Furthermore, we will generally assume that the time bin is short enough that r_i is binary, but address the more general case in Chapter 2. Finally, we must determine how to model the stochasticity in neural responses. The measured response of real neurons display varying outputs even when presented with repetitions of the exact same stimulus. Rate codes discard all information about variability, and simply treat $\vec{r}_t$ as a vector where each each entry represents the average number spikes from the corresponding neuron over repeated presentations of $\vec{s}_t$. However, multiple studies have demonstrated that the

magnitude of the variability in individual neuron responses, as well as intrinsic correlations between the variability of different neurons, plays an important part in cortical population codes [Berkes et al., 2011, Orban et al., 2016]. In this thesis we consider the ramifications of considering a particular form of parametric probability distribution for $P(\vec{r}|\vec{s})$ (see section 1.2.2) which, among other effects, determines the precise stochastic relationship between $\vec{r}$ and $\vec{s}$. Once we have made the relevant modelling assumptions about $\vec{r}$, $\vec{s}$, and $P(\vec{r}|\vec{s})$ we may ask if there are reduced representations of $\vec{r}$ that are both succinct and useful. That is, does an organism need to retain the entire description of the population response $\vec{r}$ to perform accurate inference about the stimulus or to choose the correction response, or can information about $\vec{r}$ be discarded without loss of performance? In order to answer that question we must settle upon a metric by which to determine the performance of such representations.

1.2 Efficient Coding and Shannon Information

There are many ways to judge the quality of a neural code. If one views the role of the early sensory system as providing to higher cortical areas a representation of the stimulus that is optimal for decision making, e.g. choosing the correct motor response based upon visual input, then there are two two properties a high performing neural code should achieve. The first property is that a good code should distinguish between different stimuli. That is, the output of a neural response for two different stimuli should be separated, by some metric, as much as possible. If the number of different stimuli is large or infinite then this equates to high diversity in the set of neural responses. One choice for quantifying this diversity is the Shannon Entropy of the response distribution:

$$H(\vec{R}) = -\sum_{\vec{r}} P(\vec{r}) \log P(\vec{r}) \tag{1.2}$$

Where $P(\vec{r})$ is the probability of population response $\vec{r}$, averaged over all stimuli. The second property is that, for a fixed stimulus, the neural response should vary as little as possible upon repeated presentations of the stimulus. This variability can be summed up with conditional Shannon Entropy:

$$H(\vec{R}|\vec{S}) = \langle -\sum_{\vec{r}} P(\vec{r}|\vec{s}) \log P(\vec{r}|\vec{s}) \rangle_{\vec{s}} \tag{1.3}$$

These two quantities are in obvious tension as illustrated by two extreme pathological cases. A completely random neural code that assigned $\vec{r}$ uniformly at random without dependence on $\vec{s}$, so that $P(\vec{r}) = P(\vec{r}|\vec{s}) = |R|^{-1}$, would have maximal $H(\vec{R})$ $(= \log |R|)$ but also maximal $H(\vec{R}|\vec{S})$ $(= \log |R|)$. On the other hand, a code that deterministically assigned a single particular value of $\vec{r}$, denoted $\vec{r}^*$, to every stimulus so that $P(\vec{r}) = P(\vec{r}|\vec{s}) = \delta(\vec{r}^*)$ would have $H(\vec{r}) = H(\vec{r}|\vec{s}) = 0$. This tradeoff is most directly embodied in the Shannon Mutual Information [Cover and Thomas, 2012]:

$$I(\vec{R}, \vec{S}) = H(\vec{R}) - H(\vec{R}|\vec{S}) \tag{1.4}$$

We note that while there are a number of alternative equivalent expressions for $I(\vec{R}, \vec{S})$, we will stick mostly with the formulation in (1.4). In both of the pathological cases mentioned above $I(\vec{R}, \vec{S}) = 0$, which corresponds to statistical independence between $\vec{r}$ and $\vec{s}$, signifying an uninformative code. The mutual information has a number of useful characteristics as

a measure of the quality of a code. It is non-negative and upper-bounded when at least one of $\vec{r}$ or $\vec{s}$ have finite support. $I(\vec{R}, \vec{S})$ depends only on $P(\vec{s})$ and $P(\vec{r}|\vec{s})$, and does not require making any further assumptions about the perceptually relevant distinctions between different stimuli. Additionally, the data processing inequality [Cover and Thomas, 2012] guarantees that no representation that is just a transformation of the neural response $\vec{r}$ can carry more information about the stimulus than $\vec{r}$ itself. That is, if $\vec{v} = \vec{v}(\vec{r})$ is a function of $\vec{r}$, deterministic or stochastic, and does not depend on $\vec{s}$, then $I(\vec{V}, \vec{S}) \leq I(\vec{R}, \vec{S})$. This is a useful guarantee because it ensures that if we find a biologically feasible representation of $\vec{r}$ with other desirable properties that saturates this inequality, we don't have to worry about considering tradeoffs with another representation that carries more information but has other undesirable quantities.

Finally, $I(\vec{R}, \vec{S})$ is task-agnostic: The performance of a code does not directly depend on how much a response is correlated with the correct task response. While this may seem like a negative quality at first glance, it is actually important for lower levels of sensory systems to have task-agnostic codes as their responses may be used for many different tasks. However, we note that the addition of task-related performance (or other forms of supervision) can be integrated with information theoretic objectives using the information bottleneck framework [Schneidman et al., 2001]. There is ample evidence that sensory systems are designed to optimize the mutual information between stimuli and neural responses. In the retina, the presence of the well-document center-surround receptive fields of Retinal Ganglion Cells is predicted by maximizing the information transmitted by RGC responses under the assumption that stimuli are translation invariant and gaussianly distributed with spatial correlations

that follow a power law decay [Atick and Redlich, 1990]. When response nonlinearities are taken into account, the presence of ON and OFF pathways in the retina follows from the principle of maximizing information in the presence of changing stimulus contrast. Additionally, information maximization predicts that retinal pathways should split into subpopulations with different nonlinearity thresholds [Kastner et al., 2015], and that ON and OFF pathways should have differently sized receptive fields [Ratliff et al., 2010]. Furthermore, the localized and orientation sensitive receptive fields observed in the lower layers of the visual cortex bear striking similarity to the filters produced by applying information-maximization independent component analysis to natural images [Bell and Sejnowski, 1997].

1.2.1 Additional Desirable Properties of Informative Codes

Transmitting full information is not the only metric by which to judge a neural code because there are, in general, an infinite number of ways to construct a sufficient statistic of the population response. For discrete (e.g. binary) neural responses, random linear projections of the population response will form an isomorphism, and thus a sufficient statistic, with high probability if the random weights are drawn from a continuous distribution. Random projections have been proposed as a neural coding mechanism for the olfactory system [Zhang and Sharpee, 2016], where the stimuli are themselves represented as sparse binary vectors. However, if the stimuli are continuous and/or not sparse, then performing inference on the stimulus given a sample of the encoded neural response will be difficult. A good neural code should be easily decodable, in addition to transmitting as much information as possible. This means that the decoder should be an explicit function of the decoded representation

and be implementable in terms of feedforward operations. For instance, maximum likelihood decoding will often require solving an implicit non-linear maximization problem, an operation that is not straightforwardly mapped to neural circuitry. We note that there is a fundamental link between decodability and mutual information [Agakov and Barber, 2004]. $H(\vec{S}|\vec{R})$ can be viewed as the average reconstruction error of a decoder distribution $q(\vec{s}|\vec{r})$, when the decoder distribution is bayes-optimal, i.e. equal to the true posterior: $q(\vec{s}|\vec{r}) = P(\vec{s}|\vec{r})$. However, the true posterior is generally intractable for high-dimensional continuous stimuli, and optimizing over $q(\vec{s}|\vec{r})$ can yield very complex non-linear estimates. Additionally, the code should itself be biologically feasible. Locality sensitive hashing produces highly informative codes for continuous, high dimensional vectors, but most algorithms use data structures that are unlikely to be implemented in biological systems [Dasgupta et al., 2017]. While overly complex codes are biologically infeasible, there is a tradeoff between code complexity and information loss. Even in perhaps the simplest case where the population of neurons code for identical features, the so called redundant case, it is an open question as to how the combine the outputs of individual neurons. Previous studies have demonstrated that simply pooling separate neurons, summing their responses together, discards significant amounts of information [Osborne et al., 2008a, Reich et al., 2001]. Generalizations of pooling, such as the population vector where neural responses are linear combined according to their receptive fields [Georgopoulos et al., 1986], exhibit high correlation with relevant stimuli but we shall show they also lose information. However, in chapter 2 we shall use the theory of exponential families to derive a modified form of the population vector that is guaranteed to preserve full information under certain circumstances.

1.2.2 Exponential Families

There are a natural class of statistical models that yield sufficient statistics with useful and intuitive properties. Exponential families are conditional distributions, $P(\vec{r}|\vec{s})$, that obey a particular functional form relating $\vec{r}$ to $\vec{s}$ [Wainwright and Jordan, 2008]:

$$\begin{aligned} P(\vec{r}|\vec{s}) &= h(\vec{r}) \exp(\vec{t}(\vec{r}) \cdot \vec{\gamma}(\vec{s}) - A(\vec{s})) \\ A(\vec{s}) &\equiv \log \left(\sum_{\vec{r}} h(\vec{r}) \exp(\vec{t}(\vec{r}) \cdot \vec{\gamma}(\vec{s})) \right) \end{aligned}$$

The mappings $\vec{t}(\vec{r})$: $\mathcal{R}^N \rightarrow \mathcal{R}^{D'}$ and $\vec{\gamma}(\vec{s})$: $\mathcal{R}^D \rightarrow \mathcal{R}^{D'}$ are known as the sufficient statistic and natural parameter mapping respectively. $h(\vec{r})$ is known as the base measure, and $A(\vec{s})$ is a stimulus dependent normalizing term often denoted as the log-partition function. A single exponential family is considered to have fixed $h(\vec{r})$, $\vec{t}(\vec{r})$, and $\vec{\gamma}(\vec{s})$ with different elements of the family indexed by different values of $\vec{s}$. If $\vec{\gamma}(\vec{s}) = \vec{s}$ then the exponential family is in canonical form, though an equivalent exponential family may be defined by applying an invertible affine transformation to $\vec{s}$ and the inverse of that transformation to $\vec{t}$. Exponential families have many useful properties, which will be elaborated upon in chapters 2 and 3, but the primary property we make use of is the preservation of mutual information by the sufficient statistic. If $P(\vec{r}|\vec{s})$ is an exponential family in canonical form with sufficient statistic $\vec{t}$ then the following equality holds:

$$I(\vec{R}, \vec{S}) = I(\vec{T}, \vec{S}) \tag{1.5}$$

If $D \ll N$, then (1.5) implies a substantial dimensionality reduction in the response, which we take advantage of in chapter 3. Additionally, in this same asymptotic limit, under certain conditions, we show that $\vec{t}$ itself can be linearly decoded to produce an estimate of the direction of $\vec{s}$ without having to perform nonlinear, iterative estimation.

1.2.3 Algorithmic Obstacles for Mutual Information

Though the mutual information is considered a gold-standard measure for how effective a sensory encoding is, there are several reasons why it is not often used in practice. Evaluating the Shannon Entropy, Mutual Information, and other information-theoretic quantities requires full knowledge of the probability distributions, joint and/or marginal, of the variables involved. Unless exact expressions are available for these distributions, the probabilities have to be estimated from samples drawn from those distributions. Such estimation processes invariably lead to statistical biases in the finite-sample limit [Paninski, 2003], although various methods have been proposed to try and correct for these biases in the relatively simple case of discrete variables with known support [Treves and Panzeri, 1995, Strong et al., 1998, Nemenman et al., 2002]. For continuous variables the situation is even more difficult, as identifying a density generally requires strong parametric assumptions [Gao et al., 2018]. While there are non-parametric sample based estimators for mutual information calculations involving one or more continuous variables [Kraskov et al., 2004, Gao et al., 2015], these methods are statistically biased for high dimensional variables [Gao et al., 2018]. However, we shall return to these methods after assuming an exponential family model of response distributions. Even in the case where a model for the population response conditional distribution $P(\vec{r}|\vec{s})$

is known, which is the situation we assume in this thesis, there can still be computational difficulties in evaluating mutual information. For discrete neural responses, the number of possible response patterns grows exponentially with the population size. After marginalizing over the stimulus the resulting marginal distribution over the response patterns will not factor, in general, even if the neural responses are independent conditioned on the stimulus. Thus, the computational cost of evaluating the Shannon Entropy of the marginal distribution grows exponentially with the population size, rendering this calculation intractable for realistic populations. Once again, the generalization to continuous responses introduces additional complexities, so that the calculation is prohibitive even for lower dimensional responses patterns. Though we make use of the properties of exponential families to reduce the relevant response variable from the high dimensional population response $\vec{R}$ to the relatively low dimensional sufficient statistic $\vec{T}$, this does not necessarily make the calculation any easier. There is no guarantee that the cardinality of $\vec{T}$ will be significantly smaller than $\vec{R}$, nor that the marginal distribution over $\vec{T}$ will factor. Additionally, forming an expression for $P(\vec{t}|\vec{s})$ or $P(t_d|\vec{s})$ can be difficult, as $P(\vec{t}|\vec{s})$ is implicitly a marginalization over all $\vec{r}$ that map to the same $\vec{t}$. Even when this mapping is linear, understanding the resultant probability distribution over $\vec{T}$ from a geometric perspective is challenging. The Littlewood-Offord problem is a simplified version of this problem, and only very recently has there been any progress [Tao and Vu, 2010]. However, both because of the dimensionality reduction involved in going from $\vec{R}$ to $\vec{T}$ as well as properties of exponential families that we will elaborate upon in chapter 3, the aforementioned nonparametric estimators of [Kraskov et al., 2004, Gao et al., 2015] become vastly more efficient in this case. Thus, we can make use of these methods to estimate

the information transmitted by a neural population, sidestepping the issues of computational complexity and intractable marginalization. Additionally, in cases where even after reducing in dimension the mutual information is still intractable to estimate, we can use the properties of exponential families to drive much more tractable lower bounds on the information.

1.3 Outline of Thesis

The rest of this thesis is organized as follows. Chapter 2 introduces a broad class of models for a population of spiking neurons, shows that this model class has a sufficient statistic of fixed dimension, and compares this statistic with related quantities proposed in the neuroscience literature. Additionally we consider the question of how to decode this statistic in a biologically feasible way and show it has favorable properties in an asymptotic limit. In chapter 3 we consider the related algorithmic problem of estimating the information transmitted by members of the aforementioned model class. We introduce several decompositions of the mutual information, and provide sufficient conditions for when the mutual information can be reduced to a sum of low dimensional terms. Finally we conclude in chapter 4 with a summary of our work and ideas for future extensions.

Chapter 2: Information Transmission and Sufficient Statistics in Exponential Families

In this chapter we show that for a broad class of common models of neural populations a linear statistic of the population response can be constructed that is guaranteed to transmit as much information about the stimulus as the population response. The statistic has fixed dimension, independent of the population size, and is valid even in the presence of intrinsic interneuronal correlations. We validate the performance of this statistic on a dataset of visual neural responses. Additionally we show that under certain conditions this statistic itself can serve as a reconstruction of the stimulus.

2.4 Population Models

We begin by modelling neural responses from individual neurons as a binary variable r taking a value 1 when the neuron produces a spike and 0 otherwise. To account for response

saturation and rectification, we model the probability of a spike ($r = 1$) as a saturating function of the stimulus projection onto the neuron's receptive field $\vec{w}^{(k)}$. Specifically, we choose the logistic function in order to take advantage of the properties of exponential families described below.

$$P(r_k = 1|\vec{s}) = \frac{1}{1 + exp(-2\beta^{(k)}(x^{(k)}(\vec{s}) - \alpha^{(k)}))}, \quad x^{(k)}(\vec{s}) = \vec{w}^{(k)} \cdot \vec{s} \tag{2.6}$$

Here, vector $\vec{s}$ represents the current stimulus, $\vec{w}^{(k)}$ represents the preferred stimulus or receptive field (RF) of the kth neuron, and x is the component of the stimulus along the receptive field. The parameters $\alpha^{(k)}$ and $\beta^{(k)}$ describe, respectively, the midpoint and slope of the logistic function (Figure 2.1A). $\beta^{(k)}$ will always be assumed to be positive. As a matter of notation neurons will be indexed by the letters i, j, and k and dimensions of the stimulus or neural receptive fields will be indexed by a, b, c, and d. The RF can be thought of as a pattern that, if presented, would elicit the strongest response from the neuron. Both $\vec{s}$ and $\vec{w}^{(k)} \in \mathbb{R}^D$ and $\vec{w}^{(k)}$ is assumed to be normalized. Additionally, the rescaled variables $y_k = 2r_k - 1$ will be used repeatedly for the sake of notational brevity. The expected value and variance of y_k when r_k is distributed like (2.6) have simple forms:

$$\begin{aligned} \mu_k(\vec{s}) &\equiv \langle y_k|\vec{s}\rangle = \tanh(\beta^{(k)}(x - \alpha^{(k)})) \\ \nu_k(\vec{s}) &\equiv \langle (y_k - \mu_k(\vec{s}))^2|\vec{s}\rangle = 1 - \tanh^2(\beta^{(k)}(x - \alpha^{(k)})) = 1 - \mu_k(\vec{s})^2 \end{aligned} \tag{2.7}$$

The population response is denoted as $\vec{r} \equiv (r_1, ..., r_N)^T$ or, equivalently, $\vec{y} = (y_1, ..., y_N)^T$. We will also make frequent use of the vectors $\vec{\mu}(\vec{s})$ and $\vec{\nu}(\vec{s})$ and the following two $D \times N$ matrices:

$$\mathbf{W} \equiv \left(\vec{w}^{(1)}, ..., \vec{w}^{(N)}\right) \tag{2.8}$$

$$\tilde{\mathbf{W}} \equiv \left(\beta^{(1)}\vec{w}^{(1)}, ..., \beta^{(1)}\vec{w}^{(N)}\right) \tag{2.9}$$

We focus on two models of population response: A conditionally independent population and a population with intrinsic pairwise correlations. Though the former is a special case of the latter, for clarity of exposition we begin with the special case of conditional independence.

2.4.1 Independent Neurons

To begin, we rewrite (2.6) in fully exponential form:

$$\begin{aligned} P(r_k|\vec{s}) &= \frac{e^{y_k\beta^{(k)}(\vec{w}^{(k)}\cdot\vec{s}-\alpha^{(k)})}}{2\cosh(\beta^{(k)}(\vec{w}^{(k)}\cdot\vec{s}-\alpha^{(k)}))} \\ &= e^{h^{(k)}(r_k)+\vec{s}\cdot\vec{M}^{(k)}(r_k)-A^{(k)}(\vec{s})} \end{aligned} \tag{2.10}$$

Where we have defined the functions $h^{(k)}(r_k)$, $\vec{M}^{(k)}(r_k)$, and $A^{(k)}(\vec{s})$ for notational convenience:

$$
\begin{aligned}
h^{(k)}(r_k) &= -\beta^{(k)}\alpha^{(k)}y_k && (2.11)\\
\vec{M}^{(k)}(r_k) &= \beta^{(k)}\vec{w}^{(k)}y_k && (2.12)\\
A^{(k)}(\vec{s}) &= \ln(2\cosh(\beta^{(k)}(\vec{w}^{(k)}\cdot\vec{s}-\alpha^{(k)}))) && (2.13)
\end{aligned}
$$

In the conditionally independent model the conditional distribution of $\vec{r}$ factors over each neuron.

$$P(\vec{r}|\vec{s}) = \prod_k P(r_k|\vec{s}) \tag{2.14}$$

Given (2.10) and (2.14) we can write the full distribution in exponential form.

$$P(\vec{r}|\vec{s}) = \exp(h(\vec{r}) + \vec{s}\cdot\vec{M}(\vec{r}) - A(\vec{s})) \tag{2.15}$$

Where the functions $h(\vec{r})$, $\vec{M}(\vec{r})$ and $A(\vec{s})$ are the summations of the corresponding individual neuron functions:

$$
\begin{aligned}
h(\vec{r}) &= \sum_k h^{(k)}(r_k)\\
\vec{M}(\vec{r}) &= \sum_k \vec{M}^{(k)}(r_k)\\
A(\vec{s}) &= \sum_k A^{(k)}(\vec{s})
\end{aligned}
\tag{2.16}
$$

The function $\vec{M}(\vec{r})$ is a mapping from $\{0,1\}^N$ to a finite subset of $\mathbb{R}^d$. Additionally, it can be seen as a linear (affine) transformation of $\vec{y}$, $(\vec{r})$

$$\vec{M}(\vec{r}) = \tilde{\mathbf{W}} \cdot \vec{y} = 2\tilde{\mathbf{W}} \cdot \vec{r} - \tilde{\mathbf{W}} \cdot \vec{1}_N \tag{2.17}$$

Since the neurons are conditionally independent, $A(\vec{s})$ takes the form of a sum over neurons but in general $A(\vec{s})$ serves as a normalizing constant given the forms of $h(\vec{r})$ and $\vec{M}(\vec{r})$:

$$\begin{aligned} A(\vec{s}) &= \ln\left(\sum_{\vec{r}} \exp(h(\vec{r}) + \vec{s} \cdot \vec{M}(\vec{r}))\right) && (2.18) \\ &= \sum_k \ln\left(2\cosh\left(\beta^{(k)}(\vec{w}^{(k)} \cdot \vec{s} - \alpha^{(k)})\right)\right) && (2.19) \end{aligned}$$

When performing calculations with $A(\vec{s})$ the hyperbolic cosine may cause numerical overflow so the following alternative functional form for $A(\vec{s})$ may be preferable:

$$A(\vec{s}) = \sum_k |\beta^{(k)}(\vec{w}^{(k)} \cdot \vec{s} - \alpha^{(k)})| + \ln(1 + e^{-2|\beta^{(k)}(\vec{w}^{(k)} \cdot \vec{s} - \alpha^{(k)})|}) \tag{2.20}$$

We note that the function $g(x) = \ln(2\cosh(x))$ has been utilized previously in the context of image processing as a softened version of the standard L_1 penalty $|x|$ [Hyvärinen et al., 2009].

2.4.2 Correlated Neurons

We also consider a population response model where the neurons are intrinsically correlated such that neuronal response variability from trial to trial is correlated between neurons, as is often observed in electrophysiological studies of the cortex [Zohary et al., 1994, Huang and Lisberger, 2009]. Importantly, we require that the strength of the correlations are independent of the stimulus. This is consistent with experimental results, at least on short timescales [Kohn and Smith, 2005, Granot-Atedgi et al., 2013]. We implement this by adding a term to $h(\vec{r})$:

$$h(\vec{r}, \mathbf{J}) = \sum_{i \neq j} \mathbf{J}_{ij} y_i y_j - \sum_k y_k \beta^{(k)} \alpha^{(k)} \tag{2.21}$$

In general we make two standard assumptions on the coupling matrix $\mathbf{J}$: That $\mathbf{J}_{ij} = \mathbf{J}_{ji}$ and $\mathbf{J}_{ii} = 0$. The joint distribution on population responses is again an exponential family:

$$P(\vec{r}|\vec{s}, \mathbf{J}) = exp(h(\vec{r}, \mathbf{J}) + \vec{s} \cdot \vec{M}(\vec{r}) - A(\vec{s}, \mathbf{J})) \tag{2.22}$$

$\vec{M}(\vec{r})$ is the same as before. $A(\vec{s}, \mathbf{J})$ is defined similarly as a stimulus dependent normalizing term but in general lacks a closed-form expression similar to $A(\vec{s})$. In order to examine the effects of weak correlations between neurons, we can approximate $A(\vec{s})$ perturbatively to first order in $\mathbf{J}$ [Kardar, 2007]:

$$
\begin{aligned}
A(\vec{s},\mathbf{J}) &\approx A(\vec{s},\mathbf{J}=0) + \sum_{i\neq j}\mathbf{J}_{ij}\frac{\partial A(\vec{s},\mathbf{J})}{\partial \mathbf{J}_{ij}}\bigg|_{\mathbf{J}=0} \\
&= A(\vec{s}) + \sum_{i\neq j}\mathbf{J}_{ij}\langle y_i y_j\rangle_0 \\
&= A(\vec{s}) + \sum_{i\neq j}\mathbf{J}_{ij}\mu_i(\vec{s})\mu_j(\vec{s}) \qquad (2.23)
\end{aligned}
$$

Where we have introduced the notation that $\langle\cdot\rangle_0 = \langle\cdot\rangle_{\vec{y}|\vec{s},\mathbf{J}=0}$. Similarly, for non-zero $\mathbf{J}$, computing $P(r_k|\vec{s})$ requires marginalizing over the states of all other neurons in the population and will in general differ from (2.6). However, we can once again compute approximate the expected value of y_k to first order in $\mathbf{J}$ [Kardar, 2007]:

$$
\begin{aligned}
\langle y_k\rangle_{\vec{y}|\vec{s},\mathbf{J}} &\approx \langle y_k\rangle_0 - (\langle y_k \sum_{i,j}\mathbf{J}_{ij}y_i y_j\rangle_0 - \langle y_k\rangle\langle\sum_{i,j}\mathbf{J}_{i\neq j}y_i y_j\rangle_0) \\
&= \mu_k(\vec{s}) - 2\left(1-\mu_k^2(\vec{s})\right)\sum_{i\neq k}\mathbf{J}_{ik}\mu_i(\vec{s}) \\
&= \mu_k(\vec{s}) - 2\nu_k(\vec{s})\sum_{i\neq k}\mathbf{J}_{ik}\mu_i(\vec{s}) \qquad (2.24)
\end{aligned}
$$

Equation 2.24 can be expressed more compactly in matrix and vector notation:

$$
\begin{aligned}
\langle\vec{y}\rangle_{\vec{y}|\vec{s},\mathbf{J}} &\approx \vec{\mu}(\vec{s}) - 2\Upsilon(\vec{s})\cdot\mathbf{J}\cdot\vec{\mu}(\vec{s}) \\
&= (\mathbf{I}_N - 2\Upsilon(\vec{s})\cdot\mathbf{J})\cdot\vec{\mu}(\vec{s}) \qquad (2.25)
\end{aligned}
$$

Where we have introduced the $N\times N$ diagonal matrix $\Upsilon(\vec{s})$:

$$\Upsilon(\vec{s})_{ii} = \nu_i(\vec{s}) \tag{2.26}$$

We note that for $\mathbf{J} = 0$ the two models are equivalent, and we will henceforth omit the explicit dependence of $h(\vec{r})$ and $A(\vec{s})$ on $\mathbf{J}$.

2.5 Exponential Families and Sufficient Statistics

The population response models we consider in (2.15) and (2.22) belong to exponential families with natural parameter $\vec{s}$ and sufficient statistic $\vec{M}(\vec{r})$ [Wainwright and Jordan, 2008]. As a matter of terminology, a single exponential family is considered to have fixed values of $\mathbf{J}_{ij}$, $\{\alpha^{(k)}\}$, $\{\beta^{(k)}\}$, and $\{\vec{w}^{(k)}\}$ with different members of the family indexed by different values of $\vec{s}$.

2.5.1 Information Preservation

An important result of $P(\vec{y}|\vec{s})$ being an exponential family is that the mutual information is preserved by the sufficient statistic [Cover and Thomas, 2012]:

$$I(\vec{r}, \vec{s}) = I(\vec{M}, \vec{s}) \tag{2.27}$$

To show this directly we first define a few functions for notational convenience. Specif-

ically, we define $C(\vec{M})$ as the sum of $e^{h(\vec{r})}$ over all $\vec{r}$ that map to the same value of $\vec{M}$:

$$C(\vec{M}) = \sum_{\vec{r}} e^{h(\vec{r})} \delta(\vec{M} - \vec{M}(\vec{r})) \tag{2.28}$$

The conditional and marginal distribution of $\vec{M}$ can be expressed in terms of $C(\vec{M})$, without reference to $\vec{r}$:

$$\begin{aligned} P(\vec{M}|\vec{s}) &= C(\vec{M}) \exp(\vec{s} \cdot \vec{M} - A(\vec{s})) \\ P(\vec{M}) &= C(\vec{M}) \int P(\vec{s}) \, \exp(\vec{s} \cdot \vec{M} - A(\vec{s})) \, d\vec{s} \end{aligned}$$

We note the relationships between $P(\vec{r}|\vec{s})$, $P(\vec{r})$ and $P(\vec{M}|\vec{s})$, $P(\vec{M})$ respectively:

$$\begin{aligned} P(\vec{M}|\vec{s}) &= \sum_{\vec{r}} P(\vec{r}|\vec{s}) \delta(\vec{M} - \vec{M}(\vec{r})) \\ P(\vec{M}) &= \sum_{\vec{r}} P(\vec{r}) \delta(\vec{M} - \vec{M}(\vec{r})). \end{aligned} \tag{2.29}$$

We now have the following important identity:

$$\begin{aligned} \frac{P(\vec{r}|\vec{s})}{P(\vec{r})} &= \frac{\exp(h(\vec{r}) + \vec{s} \cdot \vec{M}(\vec{r}) - A(\vec{s}))}{\int P(\vec{s}') \, \exp(h(\vec{r}) + \vec{s}' \cdot \vec{M}(\vec{r}) - A(\vec{s}')) \, d\vec{s}'} \\ &= \frac{\exp(\vec{s} \cdot \vec{M}(\vec{r}) - A(\vec{s}))}{\int P(\vec{s}') \, \exp(\vec{s}' \cdot \vec{M}(\vec{r}) - A(\vec{s}')) \, d\vec{s}'} \\ &= \frac{C(\vec{M}) \exp(\vec{s} \cdot \vec{M}(\vec{r}) - A(\vec{s}))}{\int P(\vec{s}') \, C(\vec{M}) \exp(\vec{s}' \cdot \vec{M}(\vec{r}) - A(\vec{s}')) \, d\vec{s}'} \\ &= \frac{P(\vec{M}|\vec{s})}{P(\vec{M})}. \end{aligned} \tag{2.30}$$

This equality yields Eq. (2.27) as follows:

$$\begin{aligned} I(\vec{r},\vec{s}) &= \int P(\vec{s}) \sum_{\vec{r}} P(\vec{r}|\vec{s}) \ln\left(\frac{P(\vec{r}|\vec{s})}{P(\vec{r})}\right) \\ &= \int P(\vec{s}) \sum_{\vec{r}} P(\vec{r}|\vec{s}) \ln\left(\frac{P(\vec{M}(\vec{r})|\vec{s})}{P(\vec{M}(\vec{r}))}\right) \\ &= \int P(\vec{s}) \sum_{\vec{M}} P(\vec{M}|\vec{s}) \ln\left(\frac{P(\vec{M}|\vec{s})}{P(\vec{M})}\right) \\ &= I(\vec{M},\vec{s}). \end{aligned} \tag{2.31}$$

Another corollary of (2.30) is that the posterior distribution of $\vec{s}$ given $\vec{r}$ depends only on $\vec{M}(\vec{r})$:

$$P(\vec{s}|\vec{r}) = P(\vec{s}|\vec{M}(\vec{r})) \tag{2.32}$$

Therefore, a Bayes optimal decoder needs only to carry out the weighted summation rather than keep track of which response (out of 2^N possible) was observed. Similar sufficiency properties are known for Gaussian r_k [Ma et al., 2006, Beck et al., 2008].as well as binary population models with independent and identically distributed neurons (so that the sufficient statistic follows a binomial distribution), since these population models are also examples of exponential families. To the best of our knowledge this is the first demonstration of a population model for binary neurons that are neither independent nor identically distributed that has a sufficient statistic with dimension independent of population size.

2.5.2 Cumulants of $\vec{M}$

An important question to ask given any probabilistic model $P(\vec{r}|\vec{s})$ is whether the model is identifiable with respect to $\vec{s}$. That is, do distinct values of $\vec{s}$ produce distinct distributions? As it turns out, the answer to this question has an elegant geometric structure for exponential families and a simple sufficient condition for identifiability can be stated for the conditionally independent model. We begin by noting the connection between cumulants of $\vec{M}$ with respect to $P(\vec{r}|\vec{s})$ and derivatives of $A(\vec{s})$. In statistical physics, the cumulants of the degrees of freedom of a system at equilibrium can be calculated by adding a conjugate field to the hamiltonian, deriving the log-partition function as a function of the field, and then taking gradients of the log-partition function with respect to the field. By construction, $\vec{s}$ is exactly the conjugate field to $\vec{M}$ in an exponential family model and we thus have the following formulae for the mean and covariance of $\vec{M}$.

$$\bar{M}(\vec{s}) = \langle \vec{M} \rangle_{\vec{M}|\vec{s}} = \vec{\nabla}_{\vec{s}} A(\vec{s}) \tag{2.33}$$

$$\mathbf{V}_{a,b}(\vec{s}) = \langle (M_a - \bar{M}_a(\vec{s}))(M_b - \bar{M}_b(\vec{s})) \rangle_{\vec{M}|\vec{s}} = \frac{\partial^2 A(\vec{s})}{\partial s_b \partial s_a} \tag{2.34}$$

Since covariance matrices are always positive semi-definite we see here that $A(\vec{s})$ is a convex function. The function $\bar{M}(\vec{s})$ is sometimes referred to as the mean value mapping of the family and is a mapping from $\mathbb{R}^D$ to a convex, open subset of $\mathbb{R}^D$. We see that the covariance matrix $\mathbf{V}$ is the jacobian of $\bar{M}(\vec{s})$ and thus if $\mathbf{V}$ is positive-definite then $\bar{M}(\vec{s})$ is a diffeomorphism (one to one) and $A(\vec{s})$ is a strictly convex function. Any exponential family

can be parameterized in terms of the mean value of the sufficient statistic as opposed to in terms of the natural parameter $\vec{s}$. If the family is identifiable with respect to $\vec{s}$ then there is a one to one correspondence between the two parameterizations [Banerjee et al., 2005]. Therefore, an exponential family is identifiable if and only if the following three equivalent conditions hold:

1. $A(\vec{s})$ is a strictly convex function.

2. $\bar{M}(\vec{s})$ is a diffeomorphism.

3. $\mathbf{V}(\vec{s})$ is positive definite.

The third condition is usually the easiest to show directly, and is what we will focus on for the $\mathbf{J} = 0$ case. In this setting, $\bar{M}(\vec{s})$ and $\mathbf{V}_{a,b}(\vec{s})$ take simple forms:

$$\bar{M}(\vec{s}) = \sum_k \beta^{(k)} \vec{w}^{(k)} \tanh(\beta^{(k)}(\vec{w}^{(k)} \cdot \vec{s} - \alpha^{(k)})) \tag{2.35}$$

$$\mathbf{V}_{a,b}(\vec{s}) = \sum_k (\beta^{(k)})^2 w_a^{(k)} w_b^{(k)} \left(1 - \tanh^2(\beta^{(k)}(\vec{w}^{(k)} \cdot \vec{s} - \alpha^{(k)}))\right) \tag{2.36}$$

Again, these cumulants can be expressed more compactly in matrix-vector notation:

$$\begin{aligned} \bar{M}(\vec{s}) &= \tilde{\mathbf{W}} \cdot \vec{\mu}(\vec{s}) \\ \mathbf{V}(\vec{s}) &= \tilde{\mathbf{W}} \Upsilon(\vec{s}) \tilde{\mathbf{W}}^T \end{aligned} \tag{2.37}$$

We note that the $\Upsilon(\vec{s})$ is a diagonal matrix with strictly positive elements and is

thus positive definite for all $\vec{s}$. Thus, $\mathbf{V}(\vec{s})$ will be positive definite if and only if $\tilde{\mathbf{W}}^T$ has full rank. In addition to relating identifiability of $P(\vec{r}|\vec{s})$ to $\bar{M}(\vec{s})$, we can make a stronger statement about the behavior of $\bar{M}(\vec{s})$. A vector valued mapping $\vec{g}(\vec{s})$ from $\mathbb{R}^D$ to $\mathbb{R}^D$ is called *multivariate monotone* ([Rockafellar and Wets, 2009], definition 12.1) if the following holds:

$$[\vec{g}(\vec{s}_1) - \vec{g}(\vec{s}_2)] \cdot (\vec{s}_1 - \vec{s}_2) \geq 0 \quad \forall \vec{s}_1, \vec{s}_2 \in \mathbb{R}^D \tag{2.38}$$

If strict inequality holds in (2.38) when $\vec{s}_1 \neq \vec{s}_2$ then $\vec{g}(\vec{s})$ is called *strictly multivariate monotone*. We note that this is a stronger condition than $\vec{g}(\vec{s})$ being one-to-one. Nevertheless, $\bar{M}(\vec{s})$ is always multivariate monotone, and strictly so if $P(\vec{r}|\vec{s})$ is identifiable.

As stated before $A(\vec{s})$ is intractable for nonzero $\mathbf{J}$, however we can take gradients of (2.23) to express $\bar{M}(\vec{s}, \mathbf{J})$ and $\mathbf{V}(\vec{s}, \mathbf{J})$ to first order in $\mathbf{J}$.

$$\begin{aligned} \bar{M}(\vec{s}, \mathbf{J}) &\approx \bar{M}(\vec{s}, \mathbf{J} = 0) - 2\tilde{\mathbf{W}}\Upsilon(\vec{s})\mathbf{J} \cdot \vec{\mu}(\vec{s}) \\ &= \tilde{\mathbf{W}} \left(\mathbf{I}_N - 2\Upsilon(\vec{s}) \cdot \mathbf{J}\right) \cdot \vec{\mu}(\vec{s}) \end{aligned} \tag{2.39}$$

$$\begin{aligned} \mathbf{V}(\vec{s}, \mathbf{J}) &\approx \tilde{\mathbf{W}}\Upsilon(\vec{s})\tilde{\mathbf{W}}^T - 2\tilde{\mathbf{W}}\Lambda(\vec{s})\tilde{\mathbf{W}}^T - 2\tilde{\mathbf{W}}\Upsilon(\vec{s})\mathbf{J}\Upsilon(\vec{s})\tilde{\mathbf{W}}^T \\ &= \tilde{\mathbf{W}} \left[\Upsilon(\vec{s}) - \Lambda(\vec{s}) - \Upsilon(\vec{s})\mathbf{J}\Upsilon(\vec{s})\right] \tilde{\mathbf{W}}^T \end{aligned} \tag{2.40}$$

Where the diagonal matrix $\Lambda(\vec{s})$ is defined as follows:

$$\Lambda_{ii}(\vec{s}) = -2\nu_i(\vec{s})\mu_i(\vec{s}) \sum_j J_{ij}\mu_j(\vec{s}) \tag{2.41}$$

We note that while $\mathbf{V}(\vec{s}, \mathbf{J})$ is guaranteed to be positive semi-definite, the approximate result in (2.40) is not. It is sometimes the case that the stimulus parameter is embedded in a higher dimensional space as the natural parameter so that $\vec{s} = t(\vec{z})$ where the dimension of $\vec{z}$ is less than that of $\vec{s}$ (c.f. sections 2.7 and 2.8.1). In this case the family is said to be *curved* with respect to $\vec{z}$. We note that full information transmission is preserved for curved exponential families, however when calculating the cumulants it is necessary to take gradients with respect to $\vec{s}$ and not θ.

2.6 $\vec{M}$ and The Population Vector

$\vec{M}(\vec{r})$ is closely related to the well studied "Population Vector" [Georgopoulos et al., 1986, Salinas and Abbott, 1994, Hohl et al., 2013]. For our population model, the population vector is easily expressed in terms of the y_k and $\vec{w}^{(k)}$:

$$\vec{U}(\vec{r}) = \sum_k \vec{w}^{(k)} y_k = \mathbf{W} \cdot \vec{y} \tag{2.42}$$

Similar to $\bar{M}(\vec{s})$ we can define the expected value of $\vec{U}$ conditioned on $\vec{s}$. As expected, it takes a simple form when the the components $\vec{r}$ are conditionally independent.

$$\bar{U}(\vec{s}) = \sum_k \vec{w}^{(k)} \mu_k = \mathbf{W} \cdot \vec{\mu} \tag{2.43}$$

When $D = 1$ the standard population vector reduces to a signed version of the spike count:

$$\begin{aligned} U_{count}(\vec{r}) &= \sum_k y_k = 2R_{count} - N \\ R_{count} &= \sum_k r_k \end{aligned}$$

Both U_{count} and R_{count} take $N + 1$ distinct values, whereas the scalar version of $\vec{M}$ can take as many distinct values as $\vec{r}$ (2^N) if all the β_k are distinct. We explore the relative information transmitting capabilities of $\vec{U}$ and $\vec{M}$ in a number of settings in Section 2.8.2.

2.7 Quadratic Terms

The results presented thus far can be extended in part to neural coding models with quadratic tuning of the following form:

$$\begin{aligned} P(r_k = 1|\vec{s}) &= \frac{1}{1 + \exp(-2f^{(k)}(\vec{s}))} \\ f^{(k)}(\vec{s}) &= \beta_1^{(k)}(\vec{w}^{(k)} \cdot \vec{s} - \alpha^{(k)}) + \beta_2^{(k)} \vec{s}^T \boldsymbol{\gamma}^{(k)} \vec{s} \end{aligned} \tag{2.44}$$

Where $\boldsymbol{\gamma}^{(k)}$ is now a $D \times D$ symmetric matrix representing the quadratic part of the RF.

For a population of neurons with such tuning curves we can still write the joint distribution over population responses as an exponential family with a linear sufficient statistic:

$$P(\vec{r}|\vec{s}) = \exp(h(\vec{r}) + \vec{\eta}(\vec{s}) \cdot \vec{M}(\vec{r}) - A(\vec{s})) \tag{2.45}$$

Where now the natural parameter $\vec{\eta}(\vec{s})$ and sufficient statistic $\vec{M}(\vec{r})$ separated into linear and quadratic terms:

$$\begin{aligned} \vec{\eta}(\vec{s}) &\equiv \{\vec{\eta}^{lin}(\vec{s}),\, \vec{\eta}^{quad}(\vec{s})\} \\ \vec{M}(\vec{r}) &\equiv \{\vec{M}^{lin}(\vec{r}),\, \vec{M}^{quad}(\vec{r})\} \end{aligned}$$

The linear terms are the same D-dimensional vectors as before:

$$\begin{aligned} \vec{\eta}^{lin}(\vec{s}) &= \vec{s} \\ \vec{M}^{lin}(\vec{r}) &= \sum_k \beta_1^{(k)} \vec{w}^{(k)} y_k \end{aligned}$$

The quadratic terms are $D \times D$ matrices:

$$\begin{aligned} \vec{\eta}^{quad}(\vec{s}) &= \vec{s}\vec{s}^T \\ \vec{M}^{quad}(\vec{r}) &= \sum_k \beta_2^{(k)} \boldsymbol{\gamma}^{(k)} y_k \end{aligned}$$

The notation "$\vec{\eta}(\vec{s}) \cdot \vec{M}(\vec{r})$" is understood as elementwise multiplication of all corresponding elements in $\vec{\eta}(\vec{s})$ and $\vec{M}(\vec{r})$ followed by a summation. This procedure can be generalized to any parametrized family of $f^{(k)}(\vec{s})$ that are order n polynomials of $\vec{s}$, yielding a natural parameter and sufficient statistic of dimension $\sum_{l=1}^{n} D^l$.

2.8 Numerical Simulations

In order to verify the claim that $\vec{M}$ transmits all the information about $\vec{s}$ contained in $\vec{r}$ even in cases where $\vec{U}$ did not, we carried out a series of numerical evaluations of $I(\vec{r}, \vec{s})$, $I(\vec{M}, \vec{s})$, and $I(\vec{U}, \vec{s})$. To better elucidate the effects of correlations, stimulus dimensionality, and receptive field arrangements we work with a consistent model of population coding and stimulus distribution described below.

2.8.1 Model Orientation Tuning

While the linear-nonlinear (LN) modeling framework of Eq. (2.6) is standard for describing how sensory neurons respond to stimuli, it is not a typical starting point for studies of population responses that have historically relied on tuning curves, a simplification possible when neural responses depend on one variable, such as orientation for visual or motor neurons. To facilitate easier comparison of our numerical simulations with previous literature on population coding, we will work with exponential family models that respond to an angular variable.

We note that if a neuron with an orientation sensitive receptive field, such as the one shown in Figure 2.1B), were probed by stimuli of oriented gratings with fixed contrast level

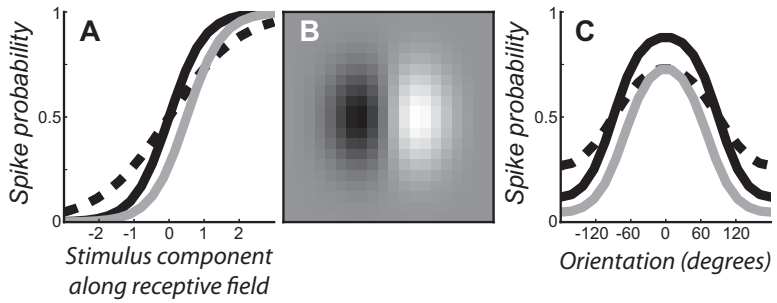

Figure 2.1: **Illustration of the correspondence between receptive field (RF) and tuning curve descriptions of the neural response.** (A) Three representative model nonlinearities that describe neural response as a logistic function of stimulus component along RF. Black and dashed lines have the same midpoints α but different β. Black and gray lines have same β but different α. (B) an example of an orientation selective receptive field. (C) Corresponding tuning curves from (A) but as a function of angle as described in the text.

then the tuning curve described in (2.6) would qualitatively display a peak firing probability at the preferred orientation. Instead of considering such a high dimensional receptive field we work with a simplified model of orientation tuning in order to provide a clearer link between the parameters $\alpha^{(k)}$, $\beta^{(k)}$ and the shape of the orientation tuning curve.

For a neuron with preferred orientation $\varphi^{(k)}$ we define $\vec{w}^{(k)} = (\cos(\varphi^{(k)}), \sin(\varphi^{(k)}))^T$ and $\vec{s}(\theta) = (\cos(\theta), \sin(\theta))^T$. Thus, in the framework of the linear-nonlinear model, the probability to observe a spike is given by:

$$P(r_k = 1|\theta) = \frac{1}{1 + e^{-2\beta^{(k)}(\cos(\varphi^{(k)} - \theta) - \alpha^{(k)})}} \tag{2.46}$$

The maximal spike rate is achieved for $\theta = \phi^{(k)}$, which is given by:

$$P_0^{(k)} = \frac{1}{1 + e^{-2\beta^{(k)}(1 - \alpha^{(k)})}} \tag{2.47}$$

The width of the orientation tuning curve, which we define as the inverse of the second

derivative of the negative logarithm of the tuning curve is:

$$\delta^{(k)} \equiv \left(-\frac{d^2}{d\theta^2} \ln \left(P(r_k = 1|\theta) \right) \Big|_{\theta=\varphi^{(k)}} \right)^{-1} = \frac{1}{2\beta^{(k)}(1 - P_0^{(k)})} \tag{2.48}$$

We can invert the above equations to express $\alpha^{(k)}$ and $\beta^{(k)}$ in terms of $P_0^{(k)}$ and $\delta^{(k)}$:

$$\begin{aligned} \alpha^{(k)} &= 1 - \delta^{(k)}(1 - P_0^{(k)}) \ln \left(\frac{P_0^{(k)}}{1 - P_0^{(k)}} \right) \\ \beta^{(k)} &= \frac{1}{2\delta^{(k)}(1 - P_0^{(k)})} \end{aligned} \tag{2.49}$$

We note that the explicit mathematical relationships given above apply only for the simplified orientation tuning model, under the assumption that the neurons are conditionally independent, and not for more general orientation selective RFs, stimuli, and correlations. We will generally specify the values of φ_k, α_k, and β_k ahead of time. In all the simulations presented θ is uniformly distributed on $[0, 2\pi]$, unless explicitly stated otherwise.

2.8.2 Results

We start by considering populations with conditionally independent neurons, so that $\mathbf{J} = 0$. The most striking demonstration of information loss from the standard population vector has been observed in cases where all of the neurons in the population have identical receptive fields [Osborne et al., 2008b]. We modelled this case by setting $\varphi_k = 0$ for all neurons. Thus the stimulus is effectively one dimensional and as stated before $\vec{M}$ and $\vec{U}$ reduce to scalars, M_{count} and U_{count} respectively. We consider two kinds of populations. In the first population, the β_k values are distributed uniformly on a $\log_{10}$ scale between 0.1

and 10.0 and the α_k values are tuned so that every neuron has $P_0 = 0.8$. The information transmission properties of this population as a function of the number of neurons are plotted in Fig 2.2A. As expected, $I(\vec{M}, \vec{s}) = I(\vec{r}, \vec{s})$ whereas $I(\vec{U}, \vec{s})$ is significantly less than $I(\vec{r}, \vec{s})$, especially for larger values of N. Because all the β_k are distinct for this population M_{count} takes 2^N distinct values, resulting in a one to one correspondence between M_{count} and $\vec{r}$. In light of this, the statement that $\vec{M}$ transmits full information may seem trivial for this case. However we can take advantage of the fact that $M_{count}, U_{count} \in \mathbb{R}$ and consider the information transmitted by binned versions these variables. Specifically we consider the smallest intervals containing the support of M_{count} and U_{count}, $[-\sum_k \beta_k, \sum_k \beta_k]$ and $[-N, N]$ respectively, and divide them uniformly into 15 bins. We then form coarse grained versions of M_{count} and U_{count}, M_{bin} and U_{bin} , corresponding to the index of the bin the variables fall into. The information transmitted by these coarse grained variables is plotted in Fig 2.2A as red and gray circles. It is notable that M_{bin} transmits close to full information (96.1 percent at $N = 10$) even when the cardinality of M_{count} is much larger than 15. We conjecture that the near optimality of M_{bin} results from the monotonicity of its mean value mapping, displayed in Fig 2.2C, such that samples drawn from $P(M_{count}|s)$ and $P(M_{count}|s')$ for very different s and s' are unlikely to fall into the same bin. This binning procedure can be generalized to vector valued $\vec{s}$ and $\vec{M}$ by first taking advantage of the multivariate monotonicity of $\bar{M}(\vec{s})$ and then choosing an appropriate vector quantization technique.

For the second population we considered a situation where the cardinality of M_{count} is considerably less than 2^N. Specifically we set $\beta_k = 1$ for all neurons and adjusted the α_k so that P_0 ranged linearly between 0.4 and 0.8. In this case there is a one to one relationship

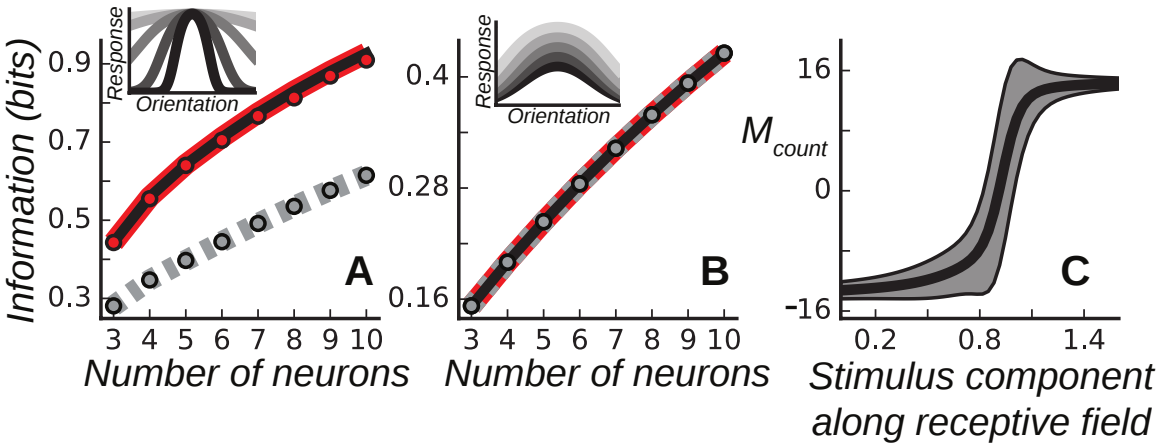

Figure 2.2: **Test of the information-preserving expression in neural populations tuned to the same stimulus.** (**A**) Information available in the full combinatorial population response (black), spike count (dashed gray), and information-preserving spike count m_{count} (red line). Circles show the results computed with binned (red) and spike count values (gray). Insets shows example tuning curves for N=5. (**B**) The standard and information-preserving spike count both capture (the curves overlap) full information when saturating nonlinearities have the same steepness, even though tuning curves have different peaks. (**C**) The expected value of the information-preserving spike count from panel (**A**) varies smoothly as a function of the stimulus component along the RF. Gray shading represents one standard deviation around the mean (black line).

between U_{count} and M_{count}. As displayed in Fig 2.2B, M_{count} still transmits full information and in this case U_{count} does as well. We note that in addition to varying P_0, the neurons also vary slightly in the curvature of their orientation response tuning curve (Fig 2.2B inset). Thus, in addition to validating that M_{count} transmits full information even in nontrivial situations, this population is an example of when the spike count of an inhomogeneous population is a sufficient statistic of the full population response. While prior work has investigated the efficient computation and optimization of information transmitted by the spike count of a neural population [McDonnell et al., 2006, McDonnel and Stocks, 2008], we believe that this work is the first demonstration of sufficient conditions for the spike count to be sufficient. For the multivariate case we also consider two different populations. For the first population the preferred orientations were distributed uniformly in $[0, \pi]$ so that $\varphi_i = \frac{i-1}{N}\pi$. $\beta_i = 1$ for all

neurons and the α_i were once again adjusted so that $P_0 = 0.8$ for all neurons. The information transmission of this populations is plotted in Fig 2.3B. Because of the effective diversity of the receptive fields in this situation there is once again a one-to-one correspondence between $\vec{M}$, $\vec{U}$ and $\vec{r}$, and all three variables transmit the same amount of information about $\vec{s}$. In the second population, half the neurons are assigned $\varphi_i = -\frac{\pi}{2}$ and the other half assigned $\varphi_i = \frac{\pi}{2}$. In each half the β_k values are distributed uniformly on a $\log_{10}$ scale between 0.1 and 10.0 and the α_k values are tuned so that every neuron has $P_0 = 0.8$, as in 2.2A. This situation, with redundancy in the receptive fields and diversity in β values, once again produces a gap between $I(\vec{M}, \vec{s})$ and $I(\vec{U}, \vec{s})$.

In Section 2.4.2 we claimed that the definition of the sufficient statistic is invariant under the presence of stimulus independent intrinsic correlations of the form defined in 2.21. To validate this claim and examine other effects of intrinsic correlations, we start by taking the populations in Fig 2.3A,B and adding correlations defined by the difference in preferred orientation:

$$\mathbf{J}_{ij} = \frac{1 + \cos(\varphi^{(i)} - \varphi^{(j)})}{10\sqrt{N}} \tag{2.50}$$

This form of $\mathbf{J}$ is consistent with the observation that correlations are stronger between neurons with similar stimulus preferences [Ecker et al., 2011, Moreno-Bote et al., 2014]. The addition of 1 in the numerator ensures that the elements of $\mathbf{J}$ are all nonnegative. The scaling by $\sqrt{N}$ is a standard practice in spin glass models to ensure that the effect of the couplings on individual neurons does not blow up for large systems. Finally we chose to additionally scale by 10 in order to distinguish the effects of weak and strong correlations, for reasons

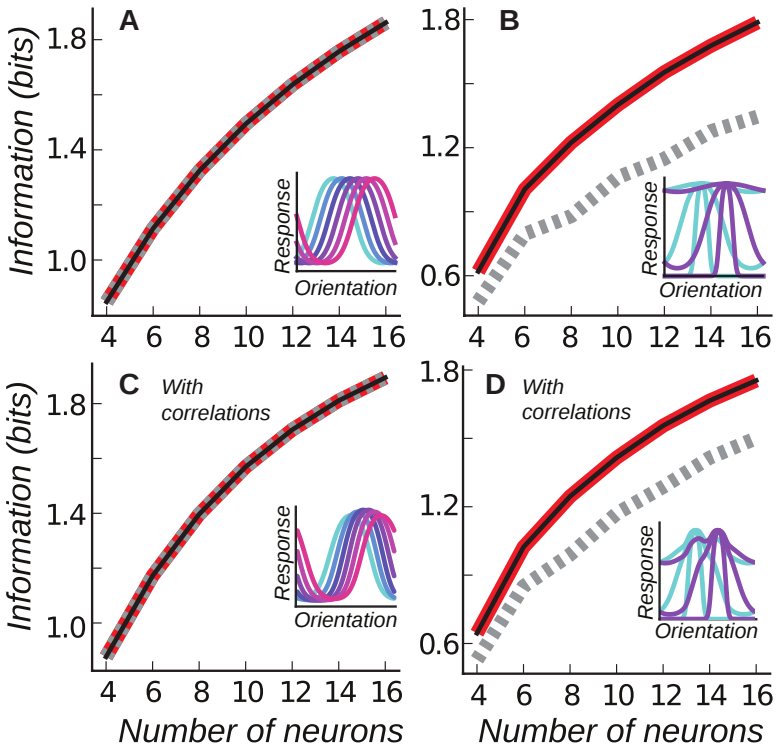

Figure 2.3: **Information-preserving vector captures full information in diverse populations and with correlated variability across neurons.** The standard population vector fails to capture full information as soon as populations contain multiple neurons with the same RF but different values of β. Insets show example population tuning curves for N=6, and $\theta \in [-2\pi, 2\pi]$. In all panels, we compare information transmitted by population response (black line) with that of information-preserving population vector (red) and standard population vector (dashed gray). Populations in panels **C** and **D** have the same parameter values as in **A**, **B**, respectively, but with the added presence of correlated variability between neurons as described in the text.

described below. All other parameters (φ_i, α_i, and β_i) remain unchanged. Fig 2.3C displays altered population from Fig 2.3A. As before, both $\vec{U}$ and $\vec{M}$ remain sufficient. The more interesting case is displayed in Fig 2.3D. Similar to Fig 2.3B there is a gap between $I(\vec{M}, \vec{s})$ and $I(\vec{U}, \vec{s})$. However the gap is significantly smaller compared to Fig 2.3B, especially for larger N, even though the presence of weak correlations has marginal effect on $I(\vec{r}, \vec{s})$.

As a more direct illustration of the effect of **J** on the difference between $I(\vec{M}, \vec{s})$ and $I(\vec{U}, \vec{s})$ we consider a simplified two neuron population where $\varphi_1 = \varphi_2 = 0$, $\beta_1 = 0.1$, $\beta_2 = 10$

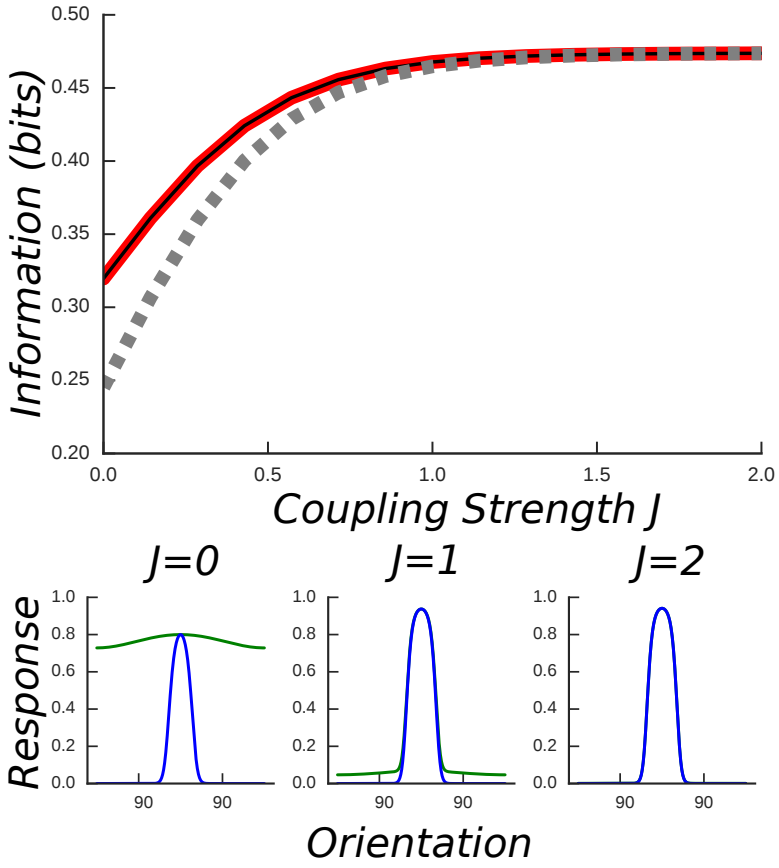

Figure 2.4: **Stronger intrinsic correlations reduce the gap between $I(\vec{M},\vec{s})$ and $I(\vec{U},\vec{s})$.** Top Panel: Information transmitted by $\vec{r}$ (black line), $\vec{M}$ (red), and $\vec{U}$ (dashed gray) for the heterogeneous and redundant two neuron population described in the text, as a function of J. Bottom Panel: The orientation tuning curves of the population at different values of J. Blue curve is for $\beta = 10$ and green curve is $\beta = 0.1$.

and the α are adjusted so that both neurons have the same peak firing rate for $\mathbf{J} = 0$. We denote this population as being *heterogenous* (different β) and *redundant* (equal φ). We define $J_{12} \equiv J$. The top panel of Fig 2.4 shows $I(\vec{r},\vec{s})$, $I(\vec{M},\vec{s})$, and $I(\vec{U},\vec{s})$ as a function of J. We note that all three values increase monotonically in J, and eventually saturate at the same value. However as soon as $J \geq 1$, $\vec{U}$ captures greater than 99% of the information transmitted by $\vec{M}$ or $\vec{r}$. Intuitively we can understand this phenomenon by first noting that since the stimulus and sufficient statistics are scalars. There are two configurations of $\vec{r}$ that U_{count} does not differentiate: $\vec{r} = (1,0)$ and $\vec{r} = (0,1)$. Both of these map to $U_{count} = 0$,

whereas $M_{count}((1,0)) = \beta_1 - \beta_2$ and $M_{count}((0,1)) = \beta_2 - \beta_1$. However, it is exactly these two configurations that become increasingly improbable for larger values of J. Understanding how strong inter-neuronal correlations affect the gap between $I(\vec{M}, \vec{s})$ and $I(\vec{U}, \vec{s})$ in more general cases is a promising direction of future work.

Another notable result displayed in Fig 2.4 is that $I(\vec{r}, \vec{s})$, $I(\vec{M}, \vec{s})$, and $I(\vec{U}, \vec{s})$ all increase as a function of J. Some previous studies, both theoretical and experimental, have reported that increasing interneuronal correlations limit the encoding performance of the population response [Zohary et al., 1994, Cohen and Maunsell, 2009]. Other studies have shown that certain forms of correlations can increase coding performance [Ecker et al., 2011], particularly for stimulus-dependent correlations [Josić et al., 2009], or that the effect can depend upon the sign of the correlations [Sompolinsky et al., 2001]. However, these studies have focused upon surrogate information measures such as Fisher Information, or the performance of maximum-likelihood or optimal-linear estimators, as opposed to the Shannon Information. Additionally most of these studies assume rate models of populations, not spiking neurons.

In order to better understand the effect of intrinsic correlations on $I(\vec{r}, \vec{s})$ we analysed two *distributed* two-neuron populations. For these populations we set $\varphi_1 = -\frac{\pi}{4}$ and $\varphi_2 = -\frac{\pi}{4}$. We considered both a *heterogeneous* and a *homogeneous* population. For the heterogeneous population we set $\beta_1 = 0.1$ and $\beta_2 = 10$. For the homogeneous population $\beta_1 = \beta_2 = 1.0$. All α were set so that $P_0 = 0.8$ when $\mathbf{J} = 0$. $\mathbf{J}$ was varied from 0 to 2 as before. As can be seen in figures 2.5 and 2.6, the heterogeneous population benefits from increased $\mathbf{J}$ while the homogeneous population does not. This result is in agreement with the results of [Ecker et al., 2011], which considered other measures of information transmission. We note that the

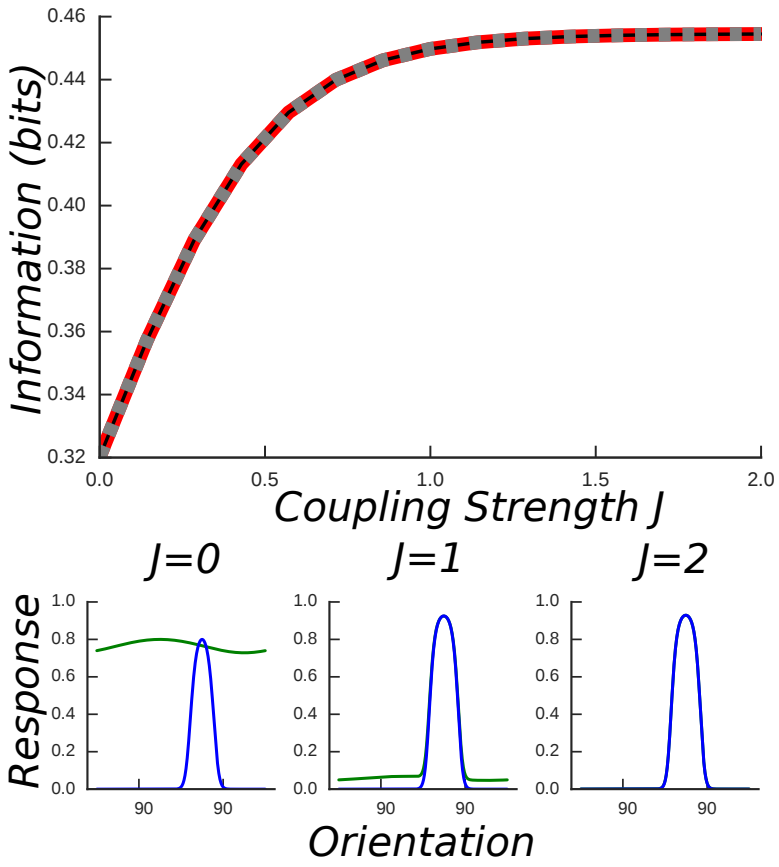

Figure 2.5: **Populations with different noise levels benefit from correlations.** Top Panel: Information transmitted by $\vec{r}$ (black line), $\vec{M}$ (red), and $\vec{U}$ (dashed gray) for the heterogeneous and distributed two neuron population described in the text, as a function of J. Bottom Panel: The orientation tuning curves of the population at different values of J. Blue curve is for $\beta = 10$ and green curve is $\beta = 0.1$.

task of determining the $\mathbf{J}$ that optimizes $I(\vec{r}, \vec{s})$ for a model of the same kind as considered here has been explored numerically for small populations in other work [Tkačik et al., 2010].

2.9 Application to V1 Data

To test if these properties of $\vec{M}$ hold for real neural populations we analyzed the responses of simultaneously recorded neurons in the primary visual cortex (V1) to natural stimuli using tetrode electrodes [Sharpee et al., 2006a] that record clusters of nearby neu-

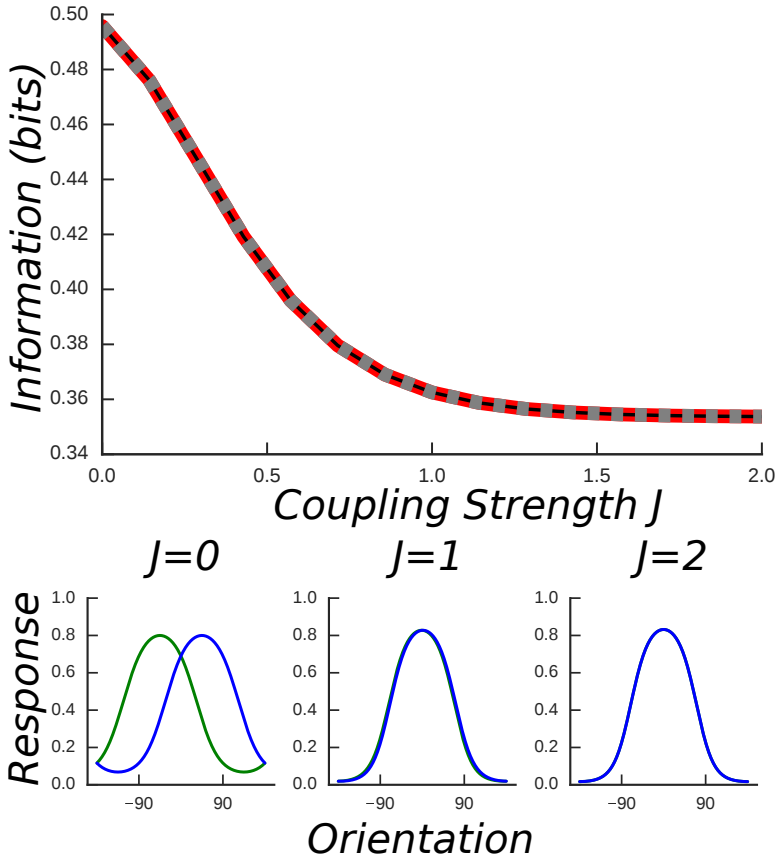

Figure 2.6: **The information transmission of homogeneous populations is diminished by correlations** Top Panel: Information transmitted by $\vec{r}$ (black line), $\vec{M}$ (red), and $\vec{U}$ (dashed gray) for the homogeneous and distributed two neuron population described in the text, as a function of J. Bottom Panel: The orientation tuning curves of the population at different values of J. Both neurons have $\beta = 1$.

rons, often with highly overlapping receptive fields. For each neuron, we estimated φ based on receptive fields and nonlinearities computed previously [Sharpee et al., 2006a]. The estimates of preferred orientation matched direct measures obtained using moving gratings [Sharpee et al., 2008b]. β values were estimated by fitting logistic functions to the neurons' firing probability as a function of stimulus projection onto $\vec{w}$. Based on these estimates, we computed the information transmitted by $\vec{r}$, $\vec{M}$, $\vec{U}$, U_{count}, and M_{count}. We only considered populations of neurons that were recorded simultaneously, and subsets therein, so that the largest population had seven neurons despite having 86 neurons in the whole dataset. For all

variables, computation of the information transmitted was done according to [Strong et al., 1998] in order to account for the finite sample estimation bias.

Reproducing previous reports [Reich et al., 2001, Osborne et al., 2008b], we found that U_{count} loses substantial information (Fig 2.7). For all populations studied the variables $\vec{M}$, $\vec{U}$, and M_{count} were in one to one correspondence with $\vec{r}$ because no pairs of neurons had exactly equal values of φ or β. Thus all three of those variables transmitted full information. However it is not always possible to measure φ, or differences in φ between neurons, to arbitrary precision. Each φ has an associated confidence interval $\Delta\varphi$ [Sharpee et al., 2006a]. We can define a measure of distinguishability between neurons i and j:

$$d_{ij} = \frac{\angle(\varphi_i, \varphi_j)}{\frac{1}{2}(\Delta\varphi_i + \Delta\varphi_j)} \tag{2.51}$$

Where the numerator is simply the angular difference between φ_i and φ_j. We consider a "worst case degeneracy" situation where the preferred orientation of pairs (or larger sets) of neurons with $d_{i,j} < 1$ are replaced by a weighted average. This was achieved with the following greedy algorithm:

1. Find the pair of neurons (or subpopulations if multiple neurons have the exact same φ) with the smallest value of d_{ij}.

2. Compute $\bar{\varphi}$, the weighted angular average of all φ for the set of neurons in step 1, where the weights are given by $\Delta\varphi_j^{-1}$. Similary compute the average value of $\Delta\varphi$.

3. For all neurons in the set found in step 1, replace φ with $\bar{\varphi}$ and $\Delta\varphi$ with its average.

4. Repeat steps 1-3 until no pair of neurons with distinct φ have $d_{ij} < 1$

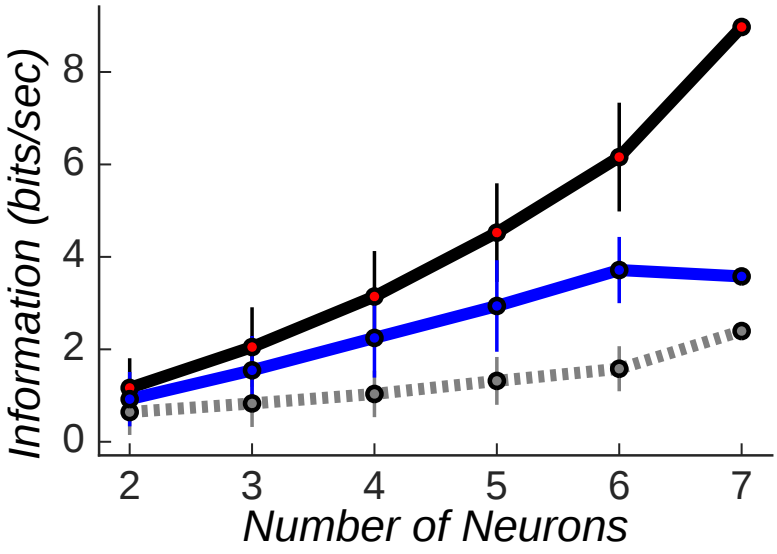

Figure 2.7: $\vec{M}$ **accounts for simultaneously recorded responses of nearby V1 neurons.** Information contained in the full population response $\vec{r}$ (solid black line) is much higher than that provided by the population spike count U_{count} (dashed gray line) and is fully captured by either $\vec{M}$, $\vec{U}$, or M_{count} (red circles, curves overlap). Blue line is information provided by $\tilde{U}$, the population vector computed based on the worst case distinguishability of preferred orientations as described in the text. Error bars are standard deviations across all sub populations of the given size.

We thus computed coarse grained versions of $\vec{M}$ and $\vec{U}$, $\tilde{M}$ and $\tilde{U}$ respectively, using the reduced set of preferred orientations produced by the above algorithm. Reducing the resolution with which the orientations of different neurons are distinguished causes the $\tilde{U}$ to lose information, but $\tilde{M}$ nevertheless transmits full information. This stability even in the worst case of angular resolution illustrates one of the practical advantages of using $\vec{M}$ over $\vec{U}$.

Estimating Mutual Information from V1 Recordings

We represent the responses of a set of N simultaneously recorded V1 cells to a stimulus binned into T segments and repeated K times as a tensor D of shape (T, K, N). T is typically 330, corresponding to time bins of 30 milliseconds. D_{tij} represents the number of times neuron

j fired in response to stimulus t on repeat i, and can be any nonnegative integer.

Converting Data to Binary Words

Since the definitions of $\{r_j\}$, $\vec{M}$, and $\vec{U}$ assume the r_j are binary we must convert D into a binary form. Let ν_{max} be the largest value in D; the maximum number of spikes over all neurons, stimuli segments, and repeats. We form a binary tensor $\tilde{D}$ of shape ($\nu_{max} \times T$,K,N) by resampling each slice $D_{t\cdot\cdot}$ into a sub tensor $\tilde{D}^{(t)}$ of shape (ν_{max},K, N) according to the following algorithm:

1. For a given value of i and j we let $n = D_{tij}$. We sample without replacement a set $L = \{\tau_l\}_{l=1}^{n}$ of n indices from the integers $\{1, ..., \nu_{max}\}$
2. We set $\tilde{D}^{(t)}_{\tau ij}$ to 1 if $\tau \in L$, and to 0 if not.
3. Steps 1 and 2 are repeated for all i and j.

After all the time slices of D are resampled, the set $\{\tilde{D}^{(t)}\}$ of binary tensors are concatenated to form a binary data tensor $\tilde{D}$ of shape ($\nu_{max} \cdot T$,K, N). Each row of $\tilde{D}_{ti\cdot}$ corresponds to a sample of $\{r_j\}$. We note that the samples described by $\tilde{D}$ correspond to time bins of length $30/\nu_{max}$ milliseconds.

Fitting Tuning Parameters and Computing $\vec{U}$ and $\vec{M}$

In order to transform samples of the population response $\{r_j\}$ into samples we need estimates of $\vec{w}_k$ and β_k for every neuron. For the analysis in figure 3 we assume that $\vec{w}_k = (\cos(\varphi_k), \sin(\varphi_k))^T$. φ_k are preferred orientations computed in [Sharpee et al., 2006a, Sharpee

et al., 2008a], along with their standard deviations $\Delta\varphi_k$. Additionally, in figure 2.7 we plot the information computed under a coarse grained realization of orientation values φ_k to take into account experimental errorbars $\Delta\phi_k$ accociated with them. The coarse graining is based on the following measure of distinguishability between orientation values for neurons i and j:

$$d_{ij} = \frac{\angle(\varphi_i, \varphi_j)}{\frac{1}{2}(\Delta\varphi_i + \Delta\varphi_j)} \tag{2.52}$$

The coarse grained realization is such that all pairs of neurons with $d_{ij} < 1$ are assigned the same value of φ. This is not strictly achievable for all sets of neurons since, for example, one neuron in a set of three may overlap in orientation with the two other neurons but the other two neurons do not. Thus, the reduced realization was approximated using the following greedy algorithm:

1. Find the pair of neurons (or subpopulations if multiple neurons have the exact same φ) with the smallest value of d_{ij}.

2. Compute $\bar{\varphi}$ as the weighted angular average of all φ for the set of neurons in step 1, with weights are given by $\Delta\varphi_j^{-1}$. Similary compute the average value of $\Delta\varphi$.

3. For all neurons in the set found in step 1, replace φ with $\bar{\varphi}$ and $\Delta\varphi$ with its average.

4. Repeat steps 1-3 until no pair of neurons with distinct φ have $d_{ij} < 1$

We use this reduced set of φ_k to compute $\vec{w}_k$ for the blue and red lines in figure 2.7. In order to estimate β_k, we fit the response rate of the kth neuron evoked by stimulus

$\vec{s}$ (averaged across the repeated presentation of this stimulus) using a logistic function:

$$r_k(\vec{s}) = \frac{r_{max}}{1 + e^{-2(\beta_k(x-\alpha_k))}} \quad x = \vec{v}_k \cdot \vec{s} \tag{2.53}$$

In this expression $\vec{v}_k$ is estimated as the maximally informative dimension for the neuron[Sharpee et al., 2006a], distinct from the $\vec{w}_k$ defined above. The parameters r_{max}, α_k, and β_k are fit by minimizing the mean square error between $r_k(\vec{s})$ and experimentally measured firing rates. We use the value of β_k fit according to this procedure in conjunction with $\vec{w}_k$ to compute the mapping between $\{r_k\}$ and $\vec{M}$, $\vec{U}$.

We note that because $\{r_k\}$ is a discrete variable with finite cardinality and the mapping between $\{r_k\}$ and $\vec{M}$ or $\vec{U}$ is deterministic, $\vec{M}$ and $\vec{U}$ are also of finite cardinality. Thus, a data tensor $\tilde{D}$ of samples of $\{r_k\}$ can be mapped to a data tensor of samples of $\vec{M}$ or $\vec{U}$ deterministically.

Adjusting for finite sample effects

We now describe how we estimate the information transmitted by $\{r_k\}$ given a data tensor $\tilde{D}$ of samples. The process for estimating the information transmitted by $\vec{M}$ and $\vec{U}$ is analogous, as they are also discrete variables of known cardinality. Our information estimate is the finite sample approximation of Shannon's Mutual Information

$$\hat{I}(\tilde{D}) = -\sum_{\{r_k\}} \hat{P}(\{r_k\}) \log_2 \hat{P}(\{r_k\}) - \frac{1}{T} \sum_t \left(-\sum_{\{r_k\}} \hat{P}_t(\{r_k\}) \log_2 \hat{P}_t(\{r_k\}) \right) \tag{2.54}$$

$\hat{P}_t(\{r_k\})$ is the empirical probability of the population response equalling $\{r_k\}$ at time bin t, computed across repeats. The marginal distribution $\hat{P}(\{r_k\})$ is simply the average of $\hat{P}_t(\{r_k\})$ across time bins:

$$\hat{P}(\{r_k\}) = \frac{1}{T}\sum_t \hat{P}_t(\{r_k\}) \tag{2.55}$$

$\hat{I}$ will be in bits, and we multiply by $\nu_{max}/0.03$ to convert $\hat{I}$ to bits per second. Since $\hat{I}$ is a biased estimate of the true mutual information for finite samples, we corrected for finite sample effects by subsampling $\tilde{D}$. We computed $\hat{I}$ using a fraction f of the repeats, for $f \in \{1.0, 0.95, 0.90, 0.85\}$, sampling repeats without replacement. We performed this subsampling ten times for each value of f. We perform linear regression on the values of $\hat{I}$ vs f^{-1} and extrapolate to $f^{-1} = 0$, the limit of infinite sample size [Strong et al., 1998]. We report the extrapolated value as our final estimate.

In figure 2.7 we include only sets of neurons recorded simultaneously, and all subsets. Thus a set of 4 simultaneously recorded neurons yields one set of size 4, 4 sets of size 3, and 6 sets of size 2.

2.10 Superiority of $\vec{M}$ over random projections

So far we have developed a statistic of the population response that is guaranteed to be sufficient as long as the population follows an exponential family distribution. Notably, this statistic transmitted full information in cases where the population vector does not. However, since the population response is inherently a discrete valued variable with finite cardinality,

it is actually quite easy to construct a linear mapping of $\vec{r}$ that will transmit full information. Consider the mapping $\hat{M}(\vec{r}) = \hat{\beta} \cdot \vec{r}$. If the components of $\hat{\beta}$ are drawn from a non-atomic distribution distribution on $\mathbb{R}$, such that the probability that $\beta_i = \beta_j$ is 0 for $i \neq j$, then $\hat{M}(\vec{r}$ will take 2^N distinct values with probability one. Thus, the random mapping $\hat{M}(\vec{r})$ is trivially information preserving. Understanding the distribution of such random projections is an open area of research in mathematics [Tao and Vu, 2010], and many compressed sensing algorithms rely upon a combination of such random projections and binning [Candès and Wakin, 2008].

In light of how easy it is to create a sufficient linear statistic simply by taking random projections, one may wonder what is useful about using the sufficient statistic $\vec{M}$. Yet information transmission is not the sole performance measure of an encoded representation. For most neural systems it is imperative that downstream neurons are able to decode useful information about the stimulus from the representation at a previous layer, and to do so in a biophysically plausible way. In the next section we turn to the complementary task of decoding the stimulus from $\vec{M}$, and show that $\vec{M}$ has certain properties that make it easily decodable or that allow it to serve as a decoder of the stimulus directly.

2.11 Treating $\vec{M}$ as a stimulus estimate

A central task for higher cortical areas involved in sensory processing is to estimate or infer the stimulus, or relevant information about the stimulus, from the activity of sensory neurons in previous layers. In the previous sections we demonstrated that the information preserving population vector $\vec{M}$ was guaranteed to transmit all the information contained

in the population response, as long as the population was well modelled by an exponential family. However, a variable that optimally *encodes* data is not guaranteed to be easily *decoded* [Schneidman et al., 2003]. The quality of a decoding is assessed by an error metric, which must be specified independently. It is not always clear what the appropriate error metric even is, particularly for continuous vector-valued variables like $\vec{s}$. In this chapter we consider the task of decoding or estimating $\vec{s}$ given $\vec{M}$ or $\vec{r}$. The mathematical field of statistical estimation and probabilistic inference is incredibly broad with numerous methods and frameworks suitable for different models and types of data. However, not all algorithms that can operate on $\vec{r}$ can be feasibly implemented in a neural circuit. We propose a method of decoding that can be implemented with simple feed forward operations.

Our decoding mechanism is a linear transform of $\vec{M}$ itself. By construction, $\vec{M}$ and $\vec{s}$ inhabit the same vector space, and changes in $\vec{s}$ are reflected in geometrically consistent changes in $\vec{M}$ (c.f. (2.38)). In light of this, we show that a linearly transformed version of $\vec{M}$, can serve as an estimate of the direction of $\vec{s}$. For large populations with $\tilde{w}$ distributed according to a multivariate gaussian we show that this estimate becomes exact, whereas a similar estimate made according to $\vec{U}$ fairs poorly. For a class of non-gaussian distributions we provide numerical evidence and analytic intuition that this estimation procedure yields a reasonable approximation.

Throughout this section we continue assume that $P(\vec{r}|\vec{s})$ is an exponential family of the same sort as described above. We make no attempt to identify the components of our proposed decoding schemes with actual specific neurobiological structures or processes, and indeed it is likely that the variables and transformations involved are at best abstractions of

more complicated underlying phenomena.

As shown above the variable $\vec{M}$ is guaranteed to capture all information about $\vec{s}$ contained in $\vec{r}$. In this section we demonstrate that under certain conditions, $\vec{M}$ also explicitly captures geometric information about $\vec{s}$. Specifically we show that a linear transformation on $\vec{M}$ yields an estimate of the orientation of $\vec{s}$. Though the information about vector amplitude is discarded, this may be acceptable for certain sensory applications such as motion discrimination.

We specifically consider the regime where $\frac{N}{D} \gg 1$ and the individual neurons are conditionally independent ($\mathbf{J} = 0$). We first define a rescaled weight vector $\tilde{w}^{(k)} = \beta^{(k)}\vec{w}^{(k)}$. We next define population-size normalized versions of $\vec{U}$ and $\vec{M}$ in terms of $\{\tilde{w}^{(k)}\}$:

$$\begin{aligned} \vec{m} &= \frac{1}{N}\sum_k \tilde{w}^{(k)}(2r^{(k)} - 1) \\ \vec{u} &= \frac{1}{N}\sum_k \frac{\tilde{w}^{(k)}}{|\tilde{w}^{(k)}|}(2r^{(k)} - 1) \end{aligned}$$

We assume that the $\tilde{w}^{(k)}$ are drawn independently from a distribution $P(\tilde{w})$. We set $\langle\tilde{w}\rangle = 0$ for simplicity, but will show to adjust for nonzero $\langle\tilde{w}\rangle$ later. The covariance matrix of $\tilde{w}$ under $P(\tilde{w})$ is denoted C and assumed to be positive definite. We also assume that probability measure of $\tilde{w} = 0$ is zero, which is the case for a smooth distribution. For now we assume that $\alpha^{(k)} = 0$ but will demonstrate how to generalize later. For fixed $\vec{s} \neq 0$, $\vec{m}$ is the sample average of $\tilde{w}(2r - 1)$ where $\tilde{w}$ and r are drawn from $P(r, \tilde{w}|\vec{s}) = P(r|\vec{s}, \tilde{w})P(\tilde{w})$. $P(r|\vec{s}, \tilde{w})$ is just the firing probability in (2.6). $\vec{u}$ can be viewed similarly. An application of

the weak law of large numbers shows that $\vec{m}$ and $\vec{u}$ converge in probability to their expected value as N grows large:

$$\begin{aligned}\vec{m} &\xrightarrow{p} \bar{m}(\vec{s}) \equiv \int P(\tilde{w})\,\tilde{w}\tanh(\tilde{w}\cdot\vec{s})\,d\tilde{w} \\ \vec{u} &\xrightarrow{p} \bar{u}(\vec{s}) \equiv \int P(\tilde{w})\,\frac{\tilde{w}}{|\tilde{w}|}\tanh(\tilde{w}\cdot\vec{s})\,d\tilde{w}\end{aligned}$$

We henceforth work with $\bar{m}(\vec{s})$ and $\bar{u}(\vec{s})$. Additionally we will also denote averaging with respect to $P(\tilde{w})$ by $\langle\cdot\rangle$. We propose the following correspondence between $\bar{m}$ and $\vec{s}$:

$$C^{-1}\bar{m}(\vec{s}) \approx \vec{s}F(\vec{s}) \tag{2.56}$$

Where $F(\vec{s})$ is a non-negative scalar function of $\vec{s}$. We note that this is not the same as the typical whitening transforms that use either $C^{-\frac{1}{2}}$ [Bell and Sejnowski, 1997], the cholesky decomposition of C, or the eigendecomposition of C [Hastie et al., 2009]. From (2.56), an estimate of the direction of $\vec{s}$ can be obtained by taking a normalized version of $C^{-1}\bar{m}(\vec{s})$ [Carandini and Heeger, 2011]. We measure the performance of this estimator using the vector correlations:

$$Corr(C^{-1}\bar{m}(\vec{s}), \vec{s}) = \frac{C^{-1}\bar{m}(\vec{s})\cdot\vec{s}}{|C^{-1}\bar{m}(\vec{s})|\,|\vec{s}|} \tag{2.57}$$

We begin by demonstrating that for gaussianly distributed $\tilde{w}$, (2.56) becomes exact and thus (2.57) goes to 1 in the large N limit. In contrast, we provide evidence that an estimator based on $\bar{u}(\vec{s})$ instead of $\bar{m}(\vec{s})$ performs poorly.

2.11.1 Gaussian $\tilde{w}$ distributions

We first state the following identity due to Stein [Stein, 1981], and later generalized in [Janzamin et al., 2014]: If $f(\tilde{w})$ is a scalar function of $\tilde{w}$ such that $\nabla_{\tilde{w}} f(\tilde{w})$ exists almost everywhere and $\langle \nabla_{\tilde{w}} f(\tilde{w}) \rangle$ also exists then the following equality in expectation holds:

$$\langle f(\tilde{w}) \nabla_{\tilde{w}} \log P(\tilde{w}) \rangle = -\langle \nabla_{\tilde{w}} f(\tilde{w}) \rangle \tag{2.58}$$

We note that generalizations exist of (2.58) to tensor valued functions $f(\tilde{w})$ and higher order derivatives [Janzamin et al., 2014]. Stein originally considered the case where $\tilde{w}$ is gaussianly distributed which we assume as well. For simplicity we assume $\tilde{w}$ to have zero mean, with the aforementioned covariance matrix C:

$$P(\tilde{w}) = \frac{1}{\sqrt{2\pi C}} e^{-\frac{1}{2}\tilde{w}^T C^{-1} \tilde{w}} \tag{2.59}$$

In this case (2.58) reduces to the following:

$$\langle \tilde{w} f(\tilde{w}) \rangle = C \langle \nabla_{\tilde{w}} f(\tilde{w}) \rangle \tag{2.60}$$

Simplified expressions for $\bar{m}(\vec{s})$ and $\bar{u}(\vec{s})$ are obtained by setting $g(\tilde{w})$ in (2.60) equal to $\tanh(\tilde{w} \cdot \vec{s})$ and $\frac{\tanh(\tilde{w} \cdot \vec{s})}{|\tilde{w}|}$ respectively:

$$\begin{aligned}
\bar{m}(\vec{s}) &= C\vec{s}\int P(\tilde{w})\,\left(1-\tanh^2(\tilde{w}\cdot\vec{s})\right)\,d\tilde{w} \\
\bar{u}(\vec{s}) &= C\vec{s}\int P(\tilde{w})\,\frac{\left(1-\tanh^2(\tilde{w}\cdot\vec{s})\right)}{|\tilde{w}|}\,d\tilde{w} - C\int P(\tilde{w})\,\frac{\tanh(\tilde{w}\cdot\vec{s})}{|\tilde{w}|^2}\frac{\tilde{w}}{|\tilde{w}|}\,d\tilde{w}
\end{aligned}$$

It is assumed that all the relevant integrals exist. Finally we multiply both $\bar{m}(\vec{s})$ and $\bar{u}(\vec{s})$ by C^{-1}:

$$C^{-1}\bar{m}(\vec{s}) = \vec{s}\int P(\tilde{w})\,\left(1-\tanh^2(\tilde{w}\cdot\vec{s})\right)\,d\tilde{w} \tag{2.61}$$

$$\begin{aligned}
C^{-1}\bar{u}(\vec{s}) &= \vec{s}\int P(\tilde{w})\,\frac{\left(1-\tanh^2(\tilde{w}\cdot\vec{s})\right)}{|\tilde{w}|}\,d\tilde{w} \\
&- \int P(\tilde{w})\,\tilde{w}\,\frac{\tanh(\tilde{w}\cdot\vec{s})}{|\tilde{w}|^3}d\tilde{w}
\end{aligned} \tag{2.62}$$

We see now that $C^{-1}\bar{m}(\vec{s})$ and the first term in $C^{-1}\bar{u}(\vec{s})$ are equal to $\vec{s}$ times an $\vec{s}$-dependent non-negative scaling factor. It is easy to show that the second term in $C^{-1}\bar{u}(\vec{s})$ is parallel to $\vec{s}$ if $P(\tilde{w})$ is a spherical distribution. That is, if $P(\tilde{w})$ depends only on $|\tilde{w}|$. Combined with the requirement that $P(\tilde{w})$ is Gaussian, this implies that $\tilde{w}$ is distributed like white noise. If $P(\tilde{w})$ is spherical then the integrand in the second term of $C^{-1}\bar{u}(\vec{s})$ is equal to $\tilde{w}$ times a function that depends only on $|\tilde{w}|$ and the angle between $\tilde{w}$ and $\vec{s}$. A simple argument based on symmetry shows that the component of $\tilde{w}$ perpendicular to $\vec{s}$ cancels out when integrating over $\mathbb{R}^D$. For non-spherical gaussian distributions, this term will in general

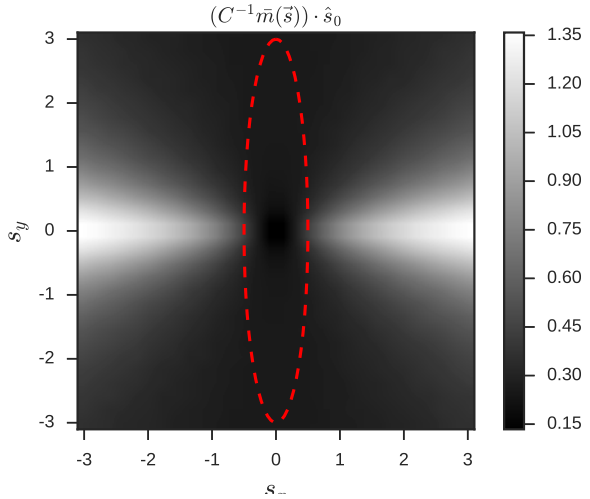

Figure 2.8: **Nonlinear compression as a function of** $\vec{s}$ Value of $C^{-1}\bar{m}(\vec{s})$ projected onto $\hat{s}_0 = \frac{\vec{s}}{|\vec{s}|}$. $P(\vec{w})$ is gaussian with $1-SD$ ellipse overlayed in red.

evaluate not to be parallel to $\vec{s}$.

Because $P(\tilde{w})$ is gaussian, we can simplify the compressive factor in $C^{-1}\bar{m}(\vec{s})$ into a one dimensional integral:

$$\begin{aligned} F(\vec{s}) &\equiv \int P(\tilde{w})\,\left(1-\tanh^2(\tilde{w}\cdot\vec{s})\right)\,d\tilde{w} \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2(\vec{s})}} e^{-\frac{x^2}{2\sigma^2(\vec{s})}}\left(1-\tanh^2(x)\right)\,dx \\ \sigma^2(\vec{s}) &= \vec{s}^T C\vec{s} \end{aligned} \tag{2.63}$$

We investigate the dependence of $F(\vec{s})$ on $\vec{s}$ in 2.8 by plotting the projection of $C^{-1}\bar{m}(\vec{s})$ onto the unit vector in the same direction as $\vec{s}$. Since we know that $C^{-1}\bar{m}(\vec{s})$ is parallel to $\vec{s}$ this figure displays how $F(\vec{s})$ compresses $C^{-1}\bar{m}(\vec{s})$ in different directions. Notably, directions of higher variance in $\vec{w}$ display higher compression.

It is possible consider more general nonlinearities $f(\cdot)$ than $\tanh(\cdot)$ as long as $f(\cdot)$

satisfies the conditions of Stein's lemma and $f'(\cdot) > 0$. We note that this does not require per se the neurons to have a monotonic response as a function of stimulus projection onto $\tilde{w}$ but rather that the nonlinearity describing the firing rate is monotonic as a function of scalar input. For example one could take an average over a distribution of thresholds $P(\alpha)$:

$$f(\tilde{w} \cdot \vec{s}) = \int P(\vec{\alpha}) \tanh(\tilde{w} \cdot \vec{s} - \alpha) \, d\alpha \tag{2.64}$$

We note that α or any other structural parameters determining the shape of $f(\cdot)$ must be distributed independently of $\tilde{w}$. To compensate for a non-zero $\langle \tilde{w} \rangle$, one can subtract from $\bar{m}(\vec{s})$ a vector of $\langle \tilde{w} \rangle \langle f(\tilde{w} \cdot \vec{s}) \rangle$, the average $\tilde{w}$ scaled by the population firing rate in response to $\vec{s}$. This procedure could be implemented using the same sorts of operations necessary to represent $\vec{M}(\vec{s})$.

2.11.2 Non-Gaussian symmetric $\tilde{w}$ distributions

In this section we examine the accuracy of (2.56) under a more general class of distributions over $\tilde{w}$: Symmetric distributions, where $P(\tilde{w}) = P(-\tilde{w})$. For simplicity we consider distributions symmetric about $\tilde{w}$ so that $\langle \tilde{w} \rangle = 0$, but all of the arguments can easily be generalized to nonzero $\langle \tilde{w} \rangle$.

We start by expanding $f(\tilde{w} \cdot \vec{s})$ to second order in $\tilde{w}$ around $\tilde{w} = 0$.

$$f(\tilde{w} \cdot \vec{s}) \approx f(0) + f'(0) \sum_i \tilde{w}_i s_i + \frac{f''(0)}{2} \sum_{ij} \tilde{w}_i \tilde{w}_j s_i s_j + O((\tilde{w} \cdot \vec{s})^3) \tag{2.65}$$

Because we have assumed that $P(\tilde{w})$ is symmetric, odd (cross) moments of $\tilde{w}$ disappear. We can thus express the i^{th} component of $\langle \tilde{w} f(\tilde{w} \cdot \vec{s}) \rangle$ to the first two leading orders in $\tilde{w}$:

$$\langle \tilde{w}_i f(\tilde{w} \cdot \vec{s}) \rangle \approx f'(0) \langle \tilde{w}_i \sum_j \tilde{w}_j \rangle s_j + \frac{f'''(0)}{3} \langle \tilde{w}_i \sum_{jkl} \tilde{w}_j \tilde{w}_k \tilde{w}_l \rangle s_j s_k s_l \tag{2.66}$$

In vector notation this reduces to a familiar form:

$$\langle \tilde{w} f(\tilde{w} \cdot \vec{s}) \rangle \approx f'(0) C \vec{s} + O(\tilde{w}^4). \tag{2.67}$$

We assume that we are in the weak coupling regime so that $|\vec{w}| < \varepsilon$ for some small ε so that we can drop terms past second order in $\tilde{w}$. Since we have assumed $f'(x) > 0$, we once again have that $C^{-1} \langle \tilde{w} f(\tilde{w} \cdot \vec{s}) \rangle \propto \vec{s}$. As an aside we note that (2.56) is trivially true when $f(x) \propto x$ because $C\vec{s} = \langle \tilde{w} \left(\tilde{w}^T \vec{s} \right) \rangle$.

2.11.3 Simulation Results

In order to validate the convergence of (2.56) for the two classes of $\tilde{w}$ distributions considered, we carried out numerical evaluations of $Corr(C^{-1} \bar{m}(\vec{s}), \vec{s})$ as a function of population size for two different model distributions. We set $D = 2$, let $P(\vec{s}) = \mathcal{N}(0, \mathbb{I}_2)$, and drew 100,000 stimulus samples. As a test of the gaussian prediction we set $P(\tilde{w}) = \mathcal{N}(0, \Sigma)$ where the covariance matrix Σ is diagonal but highly non-spherical:

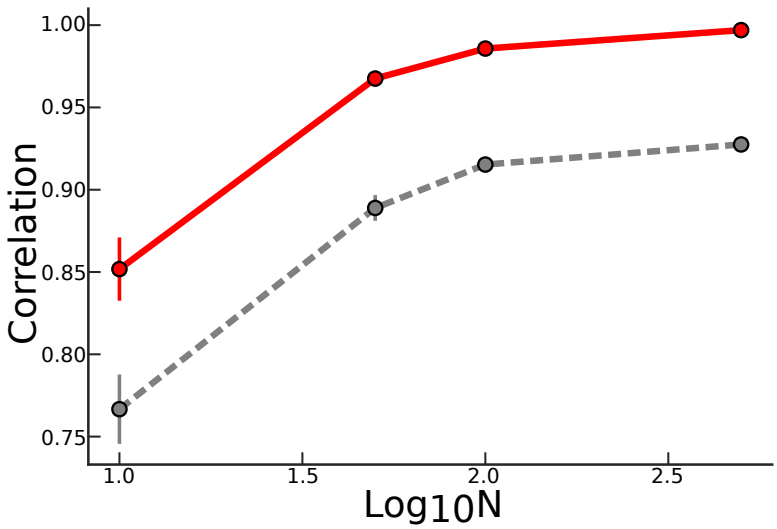

Figure 2.9: **Correlation for gaussian distributed** $\tilde{w}$ Correlation as a function of population size for $C^{-1}\bar{m}$ (red) and $C^{-1}\bar{u}$ (grey). Circles are mean across realizations of the $\tilde{w}$ and error bars are SEM across realizations.

$$\Sigma = \begin{pmatrix} 9 & 0 \\ 0 & 0.25 \end{pmatrix} \tag{2.68}$$

This distribution is the same as plotted in 2.8. For a given value of N we drew N samples from $P(\tilde{w})$, and computed $\langle \vec{w} \rangle$ and C as finite sample estimates. We then computed the average value of $Corr(C^{-1}\bar{m}(\vec{s}), \vec{s})$ and $Corr(C^{-1}\bar{u}(\vec{s}), \vec{s})$ across all samples of $\vec{s}$. We repeated this process for 50 realizations of the set of receptive fields for each N to reduce the variance due to finite sample effects, and computed both the mean across realizations and standard error of the mean across realizations. We plot the results in Fig 2.9 for $N \in \{10, 50, 100, 500\}$. The estimate based on $\bar{m}$ clearly converges to $\vec{s}$ while $\bar{u}$ performs visibly worse.

The other distribution we consider is an asymmetric uniform distribution $P(\tilde{w}) = \mathcal{U}\left(\left[-3\sqrt{3}, 3\sqrt{3}\right] \times \left[-0.5\sqrt{3}, 0.5\sqrt{3}\right]\right)$. We chose these bounds so that the population covari-

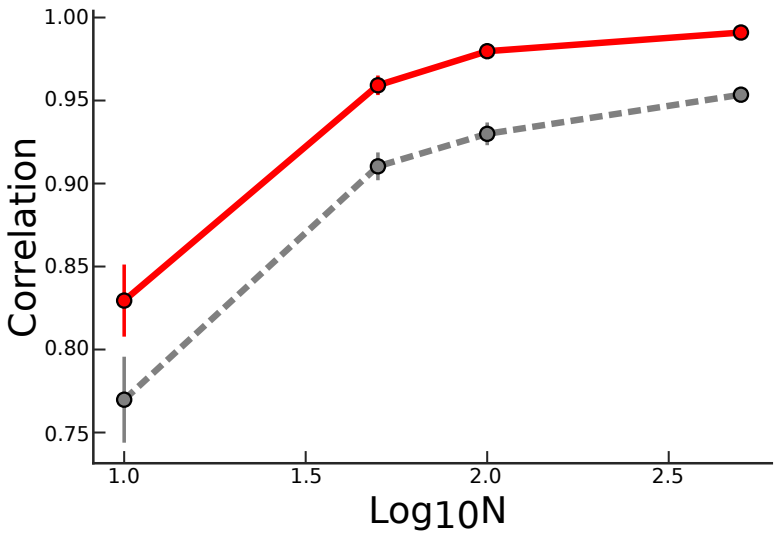

Figure 2.10: **Correlation for uniform distributed** $\tilde{w}$ Correlation as a function of population size for $C^{-1}\bar{m}$ (red) and $C^{-1}\bar{u}$ (grey). Circles are mean across realizations of the $\tilde{w}$ and error bars are SEM across realizations.

ance matrix would be the same as above. The results for this distribution are plotted in Fig 2.10. We observe similar behaviour as in the gaussian case.

2.12 Summary

In this chapter we presented a general model of a spiking neural population and demonstrated the existence of a linear sufficient statistic of bounded dimension. This statistic, the information preserving population vector $\vec{M}$, is highly interpretable as it can also serve as a reconstruction of the input stimulus under certain conditions. Yet, for many applications we may wish to explicitly compute $I(\vec{r}, \vec{s})$ and examine how different settings of the $\mathbf{W}$ affect the transmitted information in populations of realistic size. Although $\vec{M}$ can either be of lower or higher dimension than $\vec{r}$ it may not be any easier to compute $I(\vec{s}, \vec{M})$ than $I(\vec{s}, \vec{r})$ because the cardinality of the state space of $\vec{M}$ may still be quite large (and in fact will generally be

the same as that of $\vec{r}$ barring any exact symmetries in the set of receptive fields). In the next chapter we develop a more tractable estimator of $I(\vec{r}, \vec{s})$ and examine its properties.

2.13 Acknowledgments

Portions of Chapter 2 appear in a manuscript under review. John A Berkowitz, Tatyana O Sharpee. Cortical column as an information-preserving decoder of neural inputs. Under review, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 3: Quantifying information conveyed by large neuronal populations

3.14 Introduction

Information theory has the potential of answering many important questions about how neurons communicate within the brain. In particular, it can help determine whether neural responses provide sufficient amounts of information about certain stimulus features, and in this way determine whether these features could possibly affect the animal's behavior [Rieke et al., 1997, Bialek, 2012]. In addition, a number of previous studies have shown that one can understand many aspects of the neural circuit organization as those that provide maximal amounts of information under metabolic constraints [Laughlin et al., 1998, Bialek, 2012]. Key to all of these analyses is the ability to compute the Shannon mutual information [Cover and Thomas, 2012]. When estimating the information transmitted by neural populations from experimental recordings, all empirical methods produce biased estimates

[Paninski, 2003]. There are several approaches to trying to reduce or account for this bias [Nemenman et al., 2004, Strong et al., 1998, Brenner et al., 2000b, Treves and Panzeri, 1995], but these approaches do not have finite-sample guarantees and are generally ineffective when the population response is high dimensional. In order to make progress on this problem, we consider the case where the response functions of individual neurons can be measured and where the stimulus-conditional ("noise") correlations between neural responses can be described by pairwise statistics [Schneidman et al., 2006]. Historically, even with these assumptions the mutual information is notoriously difficult to compute in part due to the large number of possible responses that a set of neurons can jointly produce [Nemenman et al., 2004, Strong et al., 1998]. The number of patterns grows exponentially with both the number of time points [Strong et al., 1998, Dettner et al., 2016] and the number of neurons.

In this paper we will describe a set of approaches for computing information conveyed by responses of large neural populations. These methods build on recent advances for computing information based on linear combinations of neural responses across time [Dettner et al., 2016, Yu et al., 2010] and/or neurons [Berkowitz and Sharpee, 2018]. We will show that when each individual neuron's firing probability depends monotonically on a (potentially nonlinear) function of the stimulus, the information contained in the full population response can be completely preserved by a linear transformation of the population output. This calculation still involves computing information between high dimensional vector variables. Therefore, we further show how the full information can be effectively approximated using a sum of conditional mutual information values between pairs of low-dimensional variables. The resulting approach makes it possible to avoid the "curse of dimensionality" with

respect to the number of neurons when computing the mutual information from large neural populations.

3.15 Framework setup

Our analysis will target neural responses considered over sufficiently small time windows such that no more than one spike can be produced by any given neuron. We model the neural population as a set of binary neurons with sigmoidal tuning curves with response probability described by:

$$P(r_n = 1|\vec{s}) = \frac{1}{1 + e^{2f_n(\vec{s})}}, \tag{3.69}$$

where $\vec{s} \in \mathbb{R}^D$ is the input, $r_n \in \{-1, 1\}$ is the activity of the n^{th} neuron, and $f_n(\vec{s})$ is a scalar function of $\vec{s}$ representing the activation function of the n^{th} neuron. The population consists of N such neurons, and the population response is denoted as $\vec{r} = (r_1, ..., r_N)$. For clarity of the derivation, we will initially assume that neural responses are independent conditioned on $\vec{s}$:

$$P(\vec{r}|\vec{s}) = \prod_n P(r_n|\vec{s}), \tag{3.70}$$

and later discuss under what conditions our results generalize to the case where neural responses are correlated for a given stimulus $\vec{s}$. A few lines of algebra suffice to show that Eq. (3.70) can be expressed in the following form:

$$\begin{aligned} P(\vec{r}|\vec{s}) &= \exp\left(\sum_n r_n f_n(\vec{s}) - A_n(\vec{s})\right), \\ A_n(\vec{s}) &= \log(2\cosh(f_n(\vec{s})). \end{aligned} \tag{3.71}$$

This formulation will assist all of the approaches described below for computing the mutual information.

3.16 An unbiased estimator of information for large neural populations

In order to test the approaches described in subsequent sections, we first developed a Monte-Carlo method for computing the "ground-truth" mutual information that works for large neural populations. The approach relies on the knowledge of neural response parameters $\{f_n(\vec{s})\}$ to produce unbiased estimates of mutual information between $\vec{R}$ and $\vec{S}$ for different choices of $\{f_n(\vec{s})\}$ or $P(\vec{s})$. Here and in what follows, upper case letters (e.g. $\vec{S}$) represent random variables, while lower case letters (e.g. $\vec{s}$) represent specific values of the associated random variables. The input distribution $P(\vec{s})$ is defined by drawing N_{stim} samples; we denote this set of samples as $\{\vec{s}_\mu\}$. Because of this approximation however, $I(\vec{R}, \vec{S})$ will be bounded above by $\log(N_{\text{stim}})$ (as will be any unbiased estimator of mutual information).

Although there are several formulations of the mutual information in terms of the entropies of $\vec{R}$ and $\vec{S}$ it serves to examine just one:

$$I(\vec{R}, \vec{S}) = H(\vec{R}) - H(\vec{R}|\vec{S}). \tag{3.72}$$

Here, $H(\vec{R})$ is the Shannon entropy of the marginal distribution of $\vec{R}$ and $H(\vec{R}|\vec{S})$ is the conditional entropy of $\vec{R}$ given $\vec{S}$. Because we intend to use this estimator as a way to test the quality of other approximations, we will only consider here the case of conditionally

independent neural responses $\vec{R}$. In this case, the noise entropy $H(\vec{R}\,|\vec{S}=\vec{S}\,)$ decomposes into a sum over neurons:

$$H(\vec{R}\,|\vec{S}=\vec{s}\,)=\sum_n \bar{r}_n(\vec{s})f_n(\vec{s})-A_n(\vec{s}), \tag{3.73}$$

where $\bar{r}_n(\vec{s}\,)$ is the expected value of R_n given $\vec{s}$:

$$\bar{r}_n(\vec{s}\,)=\tanh(f_n(\vec{s}\,)). \tag{3.74}$$

We denote $\hat{H}(\vec{R}\,|\vec{S}\,)$ as the finite sample approximation to $H(\vec{R}\,|\vec{S}\,)$:

$$\hat{H}(\vec{R}\,|\vec{S}\,)=-\frac{1}{N_{\text{stim}}}\sum_\mu\sum_n \bar{r}_n(\vec{s}_\mu)f_n(\vec{s}_\mu)-A_n(\vec{s}_\mu). \tag{3.75}$$

The conditional entropy $\hat{H}(\vec{R}\,|\vec{S}\,)$ can be evaluated in $O(N*N_{\text{stim}})$ time, not including the cost of evaluating $f_n(\vec{s}\,)$. However, the marginal distribution of $\vec{r}$ will in general not factor. Thus evaluating $H(\vec{R}\,)$ requires computing the marginal $P(\vec{r}\,)$ for all $\vec{r}\in\{-1,1\}^N$. This computation grows like $O(N*N_{\text{stim}}*2^N)$. Thus, evaluation of Eq. (3.72) is known to become intractable for realistic population sizes. To derive our estimator, we begin by rewriting $H(\vec{R}\,)$:

$$\begin{aligned} H(\vec{r}\,) &= -\sum_{\vec{r}} P(\vec{r}\,)\log(P(\vec{r}\,)) = -\langle F(\vec{r}\,)\rangle_{\vec{r}}, \\ F(\vec{r}\,) &= \log\left(\langle P(\vec{r}\,|\vec{s}\,)\rangle_{\vec{s}}\right). \end{aligned} \tag{3.76}$$

We approximate the log-marginal $F(\vec{r})$ with an empirical average:

$$\hat{F}(\vec{r}) = \log\left(\frac{1}{N_{\text{stim}}}\sum_{\mu} P(\vec{r}\,|\vec{s}_\mu)\right) = \log\left(\frac{1}{N_{\text{stim}}}\sum_{\mu}\exp\left(\sum_n r_n f_n(\vec{s}_\mu) - A_n(\vec{s}_\mu)\right)\right) \quad (3.77)$$

In terms of numerical implementation, $\hat{F}(\vec{r})$ can be efficiently and stably evaluated in $O(N_{\text{stim}})$ time using the *logsumexp* function that is implemented in many numerical libraries. To approximate the averaging with respect to $P(\vec{r})$ we draw B samples of $\vec{r}$ for every $\vec{s}_\mu$, which is easily done with (3.69) and (3.70), and denote these samples as $\{\vec{r}_\nu\}$. We can thus produce an unbiased estimate of $-\langle\hat{F}(\vec{r})\rangle_{\vec{r}}$:

$$\hat{H}(\vec{R}\,) = \frac{1}{BN_{\text{stim}}}\sum_{\nu}\log\left(\frac{1}{N_{\text{stim}}}\sum_{\mu}\exp\left(\sum_n r_{n\nu} f_n(\vec{s}_\mu) - A_n(\vec{s}_\mu)\right)\right) \quad (3.78)$$

Importantly, the response entropy $\hat{H}(\vec{R}\,)$ requires $O(N * B * N_{\text{stim}}^2)$ operations, a substantial improvement over exact evaluation of $H(\vec{r})$ when $B * N_{\text{stim}} \ll 2^N$. We note that even though we are able to produce unbiased estimates of $\hat{H}(\vec{R}\,)$, this estimator systematically underestimates the "infinite sample" entropy computed with respect to $P(\vec{s})$ explicitly, i.e. not defined by input samples (see Appendix 3.20). Our Monte-Carlo estimator of $I(\vec{R}, \vec{S}\,)$ is the straightforward combination of $\hat{H}(\vec{R}\,)$ and $\hat{H}(\vec{R}\,|\vec{S}\,)$:

$$\hat{I}(\vec{R}, \vec{S}\,) = \hat{H}(\vec{R}\,) - \hat{H}(\vec{R}\,|\vec{S}\,) \quad (3.79)$$

Although $\hat{I}(\vec{R}, \vec{S}\,)$ is an unbiased estimator of the mutual information (after accounting for the approximation of $P(\vec{s})$ by samples $\{\vec{s}_\mu\}$) the variance of $\hat{H}(\vec{r})$ and thus of $\hat{I}(\vec{R}, \vec{S}\,)$ can be

difficult to quantify. However, $F(\vec{r}\,)$ is a bounded function because $\vec{r}$ has finite support (or, more generally, $F(\vec{r}\,)$ can be treated as a continuous function on the compact set $[-1,1]^N$). Thus, standard concentration bounds show that $\hat{H}(\vec{R}\,)$ is a consistent estimator of $H(\vec{R}\,)$.

In order to test our derivation that $\hat{I}(\vec{R},\vec{S}\,)$ is an unbiased estimator of $I(\vec{R},\vec{S}\,)$, we analyzed the statistics of $\hat{I}(\vec{R},\vec{S}\,)$ on a tractable neural population where $I(\vec{R},\vec{S}\,)$ can be computed exactly. We let $N = 10$ and $f_n(\vec{s}\,) = \vec{\phi}_n \cdot \vec{s}$ with $\vec{\phi}_n$ uniformly distributed along the unit circle. $P(\vec{s}\,)$ is a spherical, two-dimensional Gaussian distribution and $N_{\text{stim}} = 8,000$. We evaluate $I(\vec{R},\vec{S}\,)$ exactly, and get that $I(\vec{R},\vec{S}\,) = 1.3384$ nats. This is well below the upper bound of $\log(8,000) \approx 8.987$ nats. We computed $\hat{I}(\vec{R},\vec{S}\,)$ 100 times for $B = 1$ and $B = 3$, with $\{s_\mu\}$ fixed. For each repetition we record the residual $\hat{I}(\vec{R},\vec{S}\,) - I(\vec{R},\vec{S}\,)$. Distribution plots of the residuals are shown in Figure 3.11. For both distributions the sample mean is not significantly different from zero with $P = 0.848$ ($B = 1$) and $P = 0.851$ ($B = 3$) in a two-sided t-test. The simulation results therefore support the derivation of zero-bias in the proposed model-based Monte-Carlo estimator.

3.17 Simplifying the mutual information with sufficient statistics

3.17.1 A vector-valued sufficient statistic

The method introduced in Section 3.16 can be applied for very general formulations and parametrizations of the activation functions. However, when we constrain the activation functions to be affine we can show that $P(\vec{r}\,|\vec{s}\,)$ has especially useful properties. Specifically,

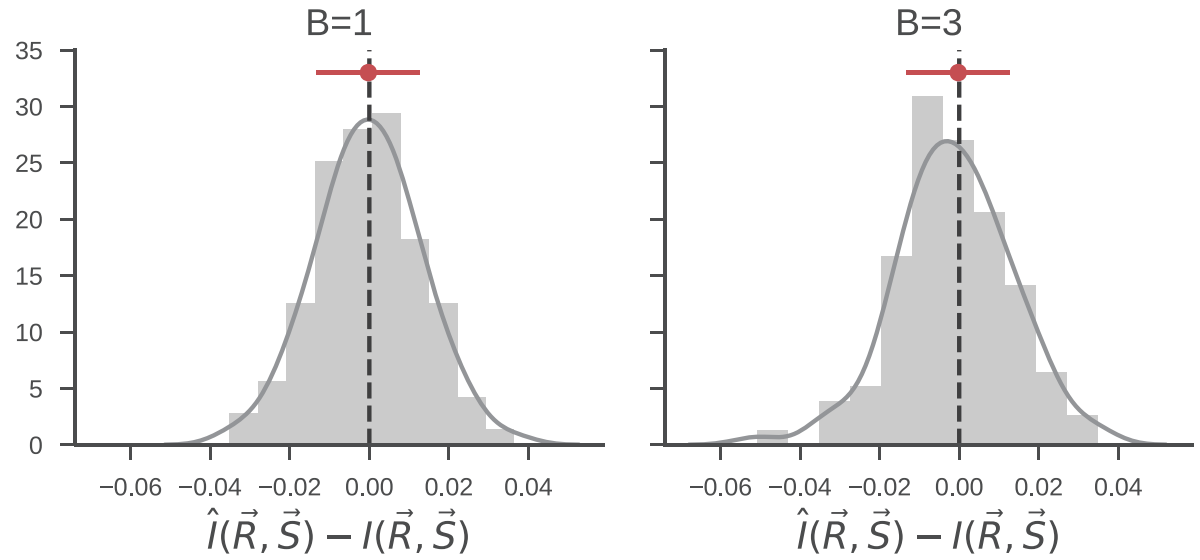

Figure 3.11: Distribution of the residuals between exact calculation and the Monte Carlo results for the test neural population described in Section 3.16. Dashed black line indicates zero, while red marker and error bar are the sample mean and standard deviation.

we assume the following parametrization of $f_n(\vec{s})$:

$$f_n(\vec{s}) = \vec{w}_n \cdot \vec{s} - \alpha_n, \; \forall n \tag{3.80}$$

While Eq. (3.80) implies a strong restriction on how stimuli drive the neural responses, some results of this section can be generalized to other activation functions. The reason for this is that even the general formulation of $P(\vec{r}|\vec{s})$ given in Eq. (3.71) can be viewed as an exponential family, with sufficient statistic $\vec{r}$ and natural parameter $\vec{f}(\vec{s})$. In particular, the framework can be extended to quadratic activation functions, which are an important model for describing neurons that are sensitive to multiple stimulus features. See Appendix 3.24 for further discussion.

If Eq. (3.80) holds, then Eq. (3.70) can be rewritten as follows:

$$P(\vec{r}|\vec{s}) = h(\vec{r}) \exp(\vec{s} \cdot \vec{t}(\vec{r}) - A(\vec{s})), \tag{3.81}$$

where,

$$\vec{t}(\vec{r}) = \mathbf{W} \cdot \vec{r}, \quad \mathbf{W} \equiv (\vec{w}_1^T, ..., \vec{w}_N^T) \tag{3.82}$$

$$h(\vec{r}) = e^{-\sum_n r_n \alpha_n}, \tag{3.83}$$

$$A(\vec{s}) = \sum_n \log\left(2\cosh(\vec{w}_n \cdot \vec{s} - \alpha_n)\right). \tag{3.84}$$

Equation (3.81) is an exponential family with sufficient statistic $\vec{t} \in \mathbb{R}^D$, natural parameter $\vec{s}$, base measure $h(\vec{r})$, and log-partition function $A(\vec{s})$ [Wainwright and Jordan, 2008].

The stimulus-conditional probability distribution $P(\vec{t}|\vec{s})$ can be defined by marginalizing over all $\vec{r}$ that map to the same $\vec{t}$:

$$\begin{aligned} P(\vec{t}|\vec{s}) &= \sum_{\vec{r}} \delta\left(\vec{t}, \vec{t}(\vec{r})\right) h(\vec{r}) \exp(\vec{s} \cdot \vec{t}(\vec{r}) - A(\vec{s})) \\ &= \exp(\vec{s} \cdot \vec{t} - A(\vec{s})) \sum_{\vec{r}} \delta\left(\vec{t}, \vec{t}(\vec{r})\right) h(\vec{r}) \\ &= \exp(\vec{s} \cdot \vec{t} - A(\vec{s})) h(\vec{t}). \end{aligned} \tag{3.85}$$

Note that $h(\vec{t}') = 0$ if there does not exist an $\vec{r}$ such that $\vec{t}' = \vec{t}(\vec{r})$. An important property of sufficient statistics is the conservation of information [Cover and Thomas, 2012]:

$$I(\vec{S}, \vec{R}) = I_{\text{vector}}(\vec{S}, \vec{T}) \tag{3.86}$$

with $\vec{T}$ defined by Eq. (3.82). Although $\vec{T}$ does not lose information relative to $\vec{r}$, it is worth making a few comments on $\vec{T}$ and Eq. (3.86). Because $\vec{R}$ is a discrete variable (with cardinality of at most 2^N) and $\vec{T}$ is a deterministic function of $\vec{R}$, then $\vec{T}$ is also a discrete variable with finite cardinality. Indeed, outside of cases of degeneracy between the columns of $\mathbf{W}$, there will generally be a one-to-one mapping between values of $\vec{R}$ and $\vec{T}$. Thus, even in cases where $D \ll N$, computing $H(\vec{T})$ can be just as difficult as computing $H(\vec{R})$. Furthermore, unlike $\vec{R}$, the components of $\vec{T}$ will generally not be conditionally independent. $H(\vec{T}\,|\vec{s})$ will thus be similarly intractable. While it may seem that we have not gained any computational advantage by transforming from $\vec{R}$ to $\vec{T}$ we will now show that Eq. (3.86) can be expressed in a convenient form that facilitates several useful approximations.

3.17.2 Decomposition of mutual information based on sufficient statistics

We start by noting that the ordering of the components of $\vec{S}$ and $\vec{T}$ is arbitrary, because applying any matching permutation to the components of $\vec{S}$ and $\vec{T}$ does not affect $I(\vec{S},\vec{R})$. We will use the following notations for components of vectors $\vec{s} = (s_1, ..., s_D)$: $s_{\neg d} = (s_1, ..., s_{d-1}, s_{d+1}, ...s_D)$, $s_{<d} = (s_1, ..., s_{d-1})$, and similarly for $s_{>d}$, $s_{\geq d}$, and $s_{\leq d}$. Note that $S_{\neg d} = (S_{<d}, S_{>d})$. Additionally, we will at times consider information theoretic quantities involving variables that are the concatenation of two other variables, such as X and Y. Such compound variables will be denoted as $\{X, Y\}$. Using these notations and applying the chain

rule for mutual information to Eq. (3.86) yields: [Cover and Thomas, 2012]:

$$I_{\text{vector}}(\vec{S}, \vec{T}\,) = \sum_{d=1}^{D} I(S_d, \vec{T}\,|S_{<d}). \tag{3.87}$$

In Eq. (3.87), $I(S_d, \vec{T}\,|S_{<d})$ is the mutual information between $\vec{T}$ and S_d conditioned on $S_{<d}$.

$$\begin{aligned} I(S_d, \vec{T}\,|S_{<d}) &= H(S_d, S_{<d}) + H(\vec{T}, S_{<d}) - H(S_d, \vec{T}, S_{<d}) - H(S_{<d}) \\ &= TC(S_{<d}, S_d, \vec{T}\,) - I(S_{<d}, S_d) - I(S_{<d}, \vec{T}\,) \\ &= I(\vec{T}, \{S_{<d}, S_d\}) - I(\vec{T}, S_{<d}) = I(\vec{T}, S_{\leq d}) - I(\vec{T}, S_{<d}) \\ &= \langle I(S_d, \vec{T}\,|s_{<d})\rangle_{s_{<d}}, \end{aligned} \tag{3.88}$$

where $TC(S_{<d}, S_d, \vec{T}\,)$ is the total correlation between $S_{<d}$, S_d, and $\vec{T}$. All four formulations of $I(S_d, \vec{T}\,|S_{<d})$ are equivalent provided they all exist. A notable situation is when there is functional dependence between S_d and $S_{<d}$, such as when the support of $S_{\leq d}$ lies on a manifold of intrinsic dimension $< d$. In this case $I(S_{<d}, S_d)$ diverges and the second line of Eq. (3.88) is ill-defined. However, it easy to show that $I(S_d, \vec{T}\,|S_{<d}) = 0$ if such a functional dependency exists using the fourth line of Eq. (3.88). Formally, let $s_d \equiv g(s_{<d})$ where $g(s_{<d}) : \mathcal{R}^{d-1} \to \mathcal{R}$ is a function defined explicitly or implicitly. Then we note that the mutual information between two variables is zero if at least one variable is constant:

$$I(S_d, \vec{T}\,|s_{<d}) = I(g(s_{<d}), \vec{T}\,|s_{<d}) = 0. \tag{3.89}$$

Thus, $I(S_d, \vec{T}\,|S_{<d})$ will also be zero:

$$I(S_d, \vec{T}\,|S_{<d}) = \langle I(S_d, \vec{T}\,|s_{<d})\rangle_{s_{<d}} = \langle I(g(s_{<d}), \vec{T}\,|s_{<d})\rangle_{s_{<d}} = 0. \tag{3.90}$$

Computing just $I(S_d, \vec{T}\,|s_{<d})$ remains challenging for the same reasons as computing $I_{\text{vector}}(\vec{S}, \vec{T}\,)$. However, we can achieve a further reduction by taking advantage of the fact that $P(\vec{t}\,|\vec{s}\,)$ is an exponential family. To see this we first express two important marginalized forms of (3.82):

$$P(\vec{t}\,|s_d, s_{<d}) = h(\vec{t}\,)\exp(s_{<d}\cdot t_{<d} + s_d t_d)\langle\exp(s_{>d}\cdot t_{>d} - A(\vec{s}\,))\rangle_{s_{>d}|s_{\leq d}}. \tag{3.91}$$

$$P(\vec{t}\,|s_{<d}) = h(\vec{t}\,)\exp(s_{<d}\cdot t_{<d})\langle\exp(s_{\geq d}\cdot t_{\geq d} - A(\vec{s}\,))\rangle_{s_{\geq d}|s_{<d}}. \tag{3.92}$$

The notation $\langle f(\vec{s}\,)\rangle_{s_{>d}|s_{\leq d}}$ denotes the expectation of $f(\vec{s}\,)$ with respect to $P(s_{>d}|s_{\leq d})$, with analogous meanings for $\langle f(\vec{s}\,)\rangle_{s_{\geq d}|s_{<d}}$ and so forth. Marginalization over conditioned variables is expressed implicitly: $P(\vec{t}\,|s_{<d}) = \langle P(\vec{t}\,|\vec{s}\,)\rangle_{s_{\geq d}}$. The important consequence of (3.91) and (3.92) is that the log-likelihood ratio of $P(\vec{t}\,|s_d, s_{<d})$ and $P(\vec{t}\,|s_{<d})$ is independent of $t_{<d}$. From this we can show that $I(S_d, \vec{T}\,|s_{<d}) = I(S_d, T_{\geq d}|s_{<d})$:

$$\begin{aligned} I(S_d, \vec{T}\,|s_{<d}) &= \left\langle \sum_{\vec{t}} P(\vec{t}\,|s_d, s_{<d}) \log\left(\frac{P(\vec{t}\,|s_d, s_{<d})}{P(\vec{t}\,|s_{<d})}\right)\right\rangle_{s_d} \\ &= \left\langle \sum_{t_{\geq d}} P(t_{\geq d}|s_d, s_{<d}) \log\left(\frac{P(t_{\geq d}|s_d, s_{<d})}{P(t_{\geq d}|s_{<d})}\right)\right\rangle_{s_d}. \\ &= I(S_d, T_{\geq d}|s_{<d}) \end{aligned} \tag{3.93}$$

This leads to our next reduction:

$$I_{\text{vector}}(\vec{S},\vec{T}\,) = \sum_{d=1}^{D} I(S_d, T_{\geq d}|S_{<d}). \tag{3.94}$$

We note that the d^{th} term in (3.87) has $D+d+1$ degrees of freedom, whereas the corresponding term in (3.94) has $D+1$ degrees of freedom. This effective dimension reduction has important algorithmic implications for the nonparametric estimators we use to compute the individual terms of (3.94) (c.f. Section 3.18). In section 3.18.1 and 3.18.2, we use Eq. (3.94) to evaluate $I_{\text{vector}}(\vec{S},\vec{T}\,)$.

3.17.3 Lower bounds for the mutual information

While Eq. (3.94) represents a significant improvement in complexity over naive evaluation of $I(\vec{S},\vec{T}\,)$, individual terms of $I_{\text{vector}}(\vec{S},\vec{T}\,)$ may be still too high dimensional to reliably evaluate. In this section, we will present a series of lower bounds on $I_{\text{vector}}(\vec{S},\vec{T}\,)$ that are more easily estimated. In particular we consider bounds that arise by replacing $T_{\geq d}$ in the d^{th} term of Eq (3.94) by a lower dimensional, deterministic transformation of $T_{\geq d}$ denoted Z_d. Applying the Data Processing Inequality (DPI) to each term in Eq (3.94), will yield a lower bound for the mutual information. There are many possible lower bounds to $I_{\text{vector}}(\vec{S},\vec{T}\,)$ of this form. We focus on a variable $Z_d = \{T_d, |T_{>d}|\}$ where $|T_{>d}|$ is the L_2-norm of $T_{>d}$. This leads to the following lower bound approximation to $I_{\text{vector}}(\vec{S},\vec{T}\,)$

$$I_{\text{iso}}(\vec{S},\vec{T}\,) = \sum_{d} I(S_d, \{T_d, |T_{>d}|\}\,|S_{<d}), \tag{3.95}$$

which we term isotropic. In Appendix 3.21, we show that this approximation becomes exact in the asymptotic limit of large neural populations, meaning that $I_{\text{iso}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\vec{S}, \vec{T})$, when the stimulus distribution is isotropic and the distribution of receptive fields (RF) $\vec{w}$ across the population is such that $A(\vec{s}) = A(|\vec{s}|)$. Notably, this is achieved when RFs uniform cover the stimulus space, meaning that $P(\vec{w})$ is described by an uncorrelated Gaussian distribution. For finite number of neurons, $A(\vec{s})$ will never be perfectly isotropic. However, for large populations ($N \gg 1$) where the receptive fields $\vec{w}$ are drawn from an isotropic distribution and the distribution of α is independent of $\vec{w}$, $A(\vec{s})$ will become isotropic asymptotically as $N \to \infty$, cf. appendix 3.21.1 for further details. The analogue of approximation Eq. (3.95) for the case where RFs and $\vec{s}$ are described by a matching correlated Gaussian distribution is described in appendix 3.21.2.

The next reduction we consider is to drop $|T_{>d}|$ from each term of Eq. (3.95):

$$I_{\text{comp-cond}}(\vec{S}, \vec{T}) = \sum_{d=1}^{D} I(S_d, T_d | S_{<d}). \tag{3.96}$$

By the data-processing inequality, it again follows that $I_{\text{iso}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T})$. Overall, one obtains a series of bounds:

$$I_{\text{vector}}(\vec{S}, \vec{T}) \geq I_{\text{iso}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T}). \tag{3.97}$$

Our final, simplest approximation is to drop the conditioning on $S_{<d}$ in each term of Eq. (3.96).

$$I_{\text{comp-ind}}(\vec{S}, \vec{T}) = \sum_{d=1}^{D} I(S_d, T_d). \tag{3.98}$$

We show in Appendix 3.22, that this last approximation becomes exact in the case where neural populations split into independent sub-populations with orthogonal RFs between sub-populations. Mathematically, this corresponds to the case where both the stimulus distribution $P(\vec{s})$ and the function $A(\vec{s})$ factor in the same basis:

$$P(\vec{s}) = \prod_{k=1}^{D} P(s'_k), \quad A(\vec{s}) = \sum_{k=1}^{D} A(s'_k). \tag{3.99}$$

In general, $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ may be greater or less than $I_{\text{comp-cond}}(\vec{S}, \vec{T})$ (or $I_{\text{vector}}(\vec{S}, \vec{T})$) [Renner and Maurer, 2002]. However when $P(\vec{s}) = \prod_d P(s_d)$, the following additional inequality holds:

$$I_{\text{comp-cond}}(\vec{S}, \vec{T}) \geq I_{\text{comp-ind}}(\vec{S}, \vec{T}). \tag{3.100}$$

To derive (3.100) we first note that we can decompose $I(S_d, \{T_d, S_{<d}\})$ (for $d > 1$) in two different ways:

$$\begin{aligned} I(S_d, \{T_d, S_{<d}\}) &= I(S_d, T_d|S_{<d}) + I(S_d, S_{<d}) \\ &= I(S_d, S_{<d}|T_d) + I(S_d, T_d) \end{aligned} \tag{3.101}$$

Equating the first and second lines of (3.101) we can rewrite the residual $I(S_d, T_d|S_{<d}) - I(S_d, T_d)$:

$$I(S_d, T_d|S_{<d}) - I(S_d, T_d) = I(S_d, S_{<d}|T_d) - I(S_d, S_{<d}) \tag{3.102}$$

Though either side of (3.102) may be positive or negative in general, when we make the assumption that $P(\vec{s})$ factors across dimension, then $I(S_d, S_{<d}) = 0$. Thus (3.102) is non-

negative and $I(S_d, T_d|S_{<d}) \geq I(S_d, T_d)$, implying (3.100).

In the opposite extreme case where the value of S_d is a deterministic function of $S_{<d}$, Eq. (3.90) can be generalized to show that $I(S_d, T_d|S_{<d}) = 0$. Thus, in this case $I(S_d, T_d) \geq I(S_d, T_d|S_{<d})$, which in turn indicates that $I_{\text{comp-ind}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T})$. For example, when the support of $P(\vec{s})$ lies on a one-dimensional curve, e.g. $\vec{S}$ represents position along a one-dimensional nonlinear track, S_d is fully determined from the values of other variables $S_{<d}$ $\forall d$ regardless of component ordering. In this case, $I_{\text{comp-ind}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T})$.

In the intermediate cases with some statistical dependencies between stimulus components, $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ is not generally guaranteed to be a lower bound to either $I_{\text{comp-cond}}(\vec{S}, \vec{T})$ or $I_{\text{vector}}(\vec{S}, \vec{T})$. Nevertheless, we observed that even for some correlated $P(\vec{s})$ $I_{\text{comp-ind}}(\vec{S}, \vec{T}) < I_{\text{comp-cond}}(\vec{S}, \vec{T})$, c.f. section 3.18.2.

3.17.4 Alternative Approximations of $I(\vec{R}, \vec{S})$

Previous authors have proposed other approximations to the mutual information. There exists a non-parametric *upper* bound to the mutual information computed in terms of pairwise relative entropies between $P(\vec{r}|\vec{s})$ and $P(\vec{r}|\vec{s}')$ [Haussler et al., 1997, Kolchinsky et al., 2017]:

$$I_{\text{k-w}}(\vec{R}, \vec{S}) = -\int d\vec{s} P(\vec{s}) \log \left(\int d\vec{s}' P(\vec{s}') \exp\left(-D_{KL}(P(\vec{r}|\vec{s})||P(\vec{r}|\vec{s}'))\right) \right) \tag{3.103}$$

The model we consider for $P(\vec{r}\,|\vec{s}\,)$ is an exponential family and thus has a tractable relative entropy [Banerjee et al., 2005]:

$$D_{KL}(P(\vec{r}\,|\vec{s}\,)||P(\vec{r}\,|\vec{s}\,') = \sum_n \left(\tanh(f_n(\vec{s}\,))f_n(\vec{s}\,) - A_n(\vec{s}\,)\right) - \left(\tanh(f_n(\vec{s}\,))f_n(\vec{s}\,') - A_n(\vec{s}\,')\right) \tag{3.104}$$

In Eq. (3.103) we have used the generalized definitions of Section 3.16. The evaluation of the upper bound (3.103) is quadratic in the sample size N_{stim} as opposed to $O(N_{\text{stim}} \log^2 N_{\text{stim}})$ for the estimator in section 3.18. In the limit where $N \gg D$, another popular approximation exist based on Fisher information [Brunel and Nadal, 1998]; it can be computed with $O(N_{\text{stim}})$ operations. Recent work has shown that this approximation is valid only for certain classes of input distributions [Huang and Zhang, 2018]. In appendix 3.23 we discuss the relationship between this approximation and $I_{\text{Fisher}}(\vec{R}, \vec{S}\,)$. We include numerical comparisons between $I_{\text{k-w}}(\vec{R}, \vec{S}\,)$ and the methods proposed in this paper in sections 3.18.1 and 3.18.2. However, we found that the Fisher Information approximation drastically overestimated the true mutual information. Therefore to avoid obscuring differences between other results, the approximation based on Fisher Information is not included in Figures 3.12-3.13. Full plots including this approximation can be found in Appendix 3.23.

We note that there are other variational approximations to mutual information [Belghazi et al., 2018, Barber and Agakov, 2003]. However, because comparing the information for different choices of the parameters of $P(\vec{r}\,|\vec{s}\,)$ and $P(\vec{s}\,)$ requires training a different variational approximation each time, direct comparison requires substantial computational resources and we leave them for future work.

3.18 Numerical Simulations

We now test the performance of the above described bounds under several representative situations that include correlated and uncorrelated stimulus distributions, and isotropic and anisotropic receptive field distributions, including experimentally recorded receptive fields from the primary visual cortex, as well as the case where intrinsic "noise" correlations are present.

To empirically estimate the bounds on mutual information [$I(S_d, T_d)$, $I(S_d, T_d|S_{<d})$, $I(S_d, \{T_d, |T_{>d}|\}\,|S_{<d})$, and $I(S_d, T_{\geq d}|S_{<d})$], we use the KSG estimator [Kraskov et al., 2004], a non-parametric method based on distributions of K nearest-neighbor distances. We chose to use the KSG estimator because even though we have reduced the mutual information between two high-dimensional variables into a sum over pairs of scalars, computing even just $I(S_d, T_d)$ can still be a daunting task, even more so for terms involving conditional informations. T_d may still have exponentially large cardinality, and complicated interdependencies between components of $\vec{T}$ present difficulties in forming explicit expressions for $P(t_d|s_d)$, so exact evaluation of $H(T_d)$ and $H(T_d|S_d)$ is not feasible at present. The KSG estimator requires only that we can draw N_{stim} samples of $\vec{S}$ and $\vec{T}$ from $P(\vec{s}, \vec{t})$, discarding the unused components. Sampling from $P(\vec{s}, \vec{t})$ is easily done given samples from $P(\vec{s})$. Given a sample $\vec{s}$, we draw $\vec{r}$ from $P(\vec{r}|\vec{s})$, which is easily done because of Eq. (3.70), and transform $\vec{r}$ into $\vec{t}$ using (3.82). This estimator has complexity $O(N_{\text{stim}} \log^2 N_{\text{stim}})$ when implemented with KD-Trees. For the case of two scalar variables the ℓ_2 error of the estimate decreases like $1/\sqrt{N_{\text{stim}}}$ [Gao et al., 2018], though if the true value of the mutual information is very high then the error may still

be large [Gao et al., 2015]. In order to partially alleviate this error, we use the PCA based Local Nonuniformity Correction of [Gao et al., 2015] (KSG-LNC). We extend this estimator to compute the conditional mutual information terms using a decomposition analogous to the second line of Eq. (3.88), and set the nonuniformity threshold hyperparameter according to the heuristics suggested in [Gao et al., 2015]. Additionally, we assume that the distribution of $\vec{S}$ (and thus $S_{\leq d}$, $S_{<d}$, and S_d $\forall d$) is non-atomic. Thus, because $\vec{T}$ is discrete but real valued, the KSG estimator is applicable as neither $\vec{T}$ nor $\vec{S}$ is a mixed continuous-discrete variable [Gao et al., 2017].

3.18.1 Large populations responding to uncorrelated stimuli

We evaluated the performance of the bounds on information developed in section 3.17.3 for large populations ranging from $N \approx 100$ to $1,000$. Specifically, to test the performance of $I_{\text{iso}}(\vec{S}, \vec{T})$ we chose a highly isotropic population and stimulus distribution. We set $D = 3$, $M = 8,000$, and let $P(\vec{s})$ be a zero mean gaussian with unit covariance matrix. For each value of N, the $\vec{w}_n$ were placed uniformly on the surface of the unit sphere, using the regular placement algorithm of [Deserno, 2004]. Because N is too large for exact evaluation of $H(\vec{r})$ ground truth values were estimated using the Monte Carlo estimator $\hat{I}(\vec{R}, \vec{S})$ of section 3.16 with $B = 3$. Results are plotted in Figure 3.12. We find that for large N, $I_{\text{iso}}(\vec{S}, \vec{T})$ tightly approximates $I_{\text{vector}}(\vec{S}, \vec{T})$ and both are accurate approximations to $\hat{I}(\vec{R}, \vec{S})$, strongly outperforming $I_{\text{k-w}}(\vec{R}, \vec{S})$. We note that for this case the upper bound of $\log(N_{\text{stim}}) = \log(8,000) \approx 9$ (nats) is well above all of the curves other than $I_{\text{k-w}}(\vec{R}, \vec{S})$, which is already known to be an upper bound to $I(\vec{R}, \vec{S})$, demonstrating that we are in the well-sampled

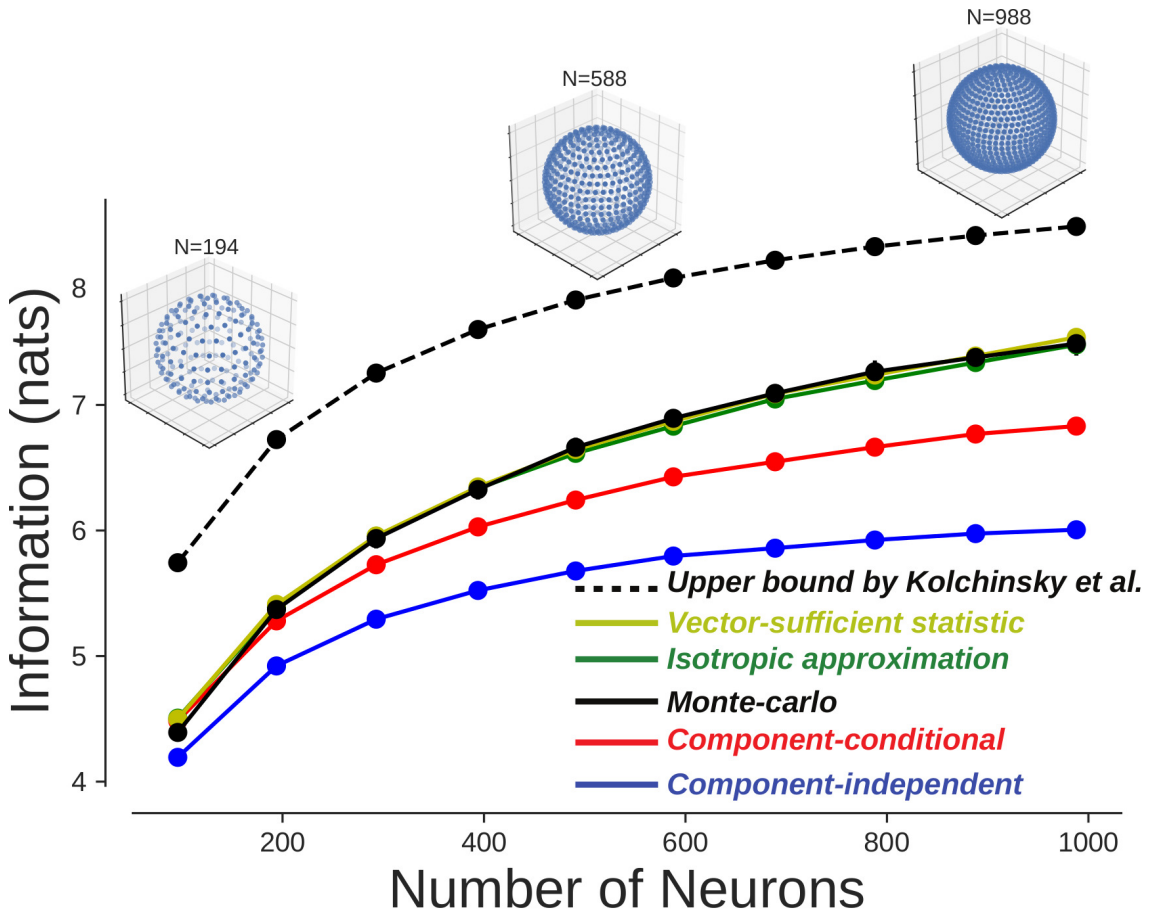

Figure 3.12: Information curves for neural populations with uncorrelated RF and stimulus distributions. Lines and error bars are mean and standard deviation over ten repeats of the estimator. Insets show RF distribution for several population sizes.

regime. Once again we see that inequalities (3.97) and (3.100) hold.

3.18.2 Correlated stimulus distributions

We now consider the case of correlated Gaussian stimuli. and model $P(\vec{s})$ as a zero-mean Gaussian with a full-rank non-diagonal covariance matrix $\mathbf{C}$. To better understand the effects of stimulus correlations we also perform computations in stimulus bases where components are independent. For this, we decompose $\mathbf{C}$ as $\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ where $\mathbf{V}$ is an

orthogonal matrix whose columns are the eigenvectors of $\mathbf{C}$.

$$\hat{S} = \mathbf{V}^T \vec{S}, \quad \hat{T} = \mathbf{V}^T \vec{T}. \tag{3.105}$$

Note that we have $\hat{t} \cdot \hat{s} = \vec{t} \cdot \vec{s}$. It is easy to see that $P(\hat{t}|\hat{s})$ is also an exponential family. Additionally because the mappings from $\vec{s}$ to $\hat{s}$ and $\vec{t}$ to $\hat{t}$ are diffeomorphisms the information is preserved:

$$I_{\text{vector}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\hat{S}, \hat{T}). \tag{3.106}$$

We note that while Eq. (3.106) holds in principle, in practice we may see variation as the KSG family of estimators is not invariant to under diffeomorphisms. Importantly we also note that Eq. (3.93) holds for $(\hat{S}, \hat{T})$. Given samples from $P(\vec{s}, \vec{t})$, we automatically have samples from $P(\hat{s}, \hat{t})$. We can straightforwardly generalize $I_{\text{vector}}(\vec{S}, \vec{T})$, $I_{\text{comp-cond}}(\vec{S}, \vec{T})$, $I_{\text{comp-ind}}(\vec{S}, \vec{T})$, and $I_{\text{iso}}(\vec{S}, \vec{T})$ to $I(\hat{S}, \hat{T})$, $I_{\text{comp-cond}}(\hat{S}, \hat{T})$, $I_{\text{comp-ind}}(\hat{S}, \hat{T})$, and $I_{\text{iso}}(\hat{S}, \hat{T})$ respectively. Eq. (3.106) does not generalize to I_{iso}, $I_{\text{comp-cond}}$, or $I_{\text{comp-ind}}$ as they are not expressible as mutual information quantities between two variables. We note that $I(\vec{R}, \vec{S}) = I(\vec{R}, \hat{S})$ so we do not modify $\hat{I}(\vec{R}, \vec{S})$.

Simulations in Figure 3.13 were done using the following stimulus covariance matrix

$$\mathbf{C} = \begin{pmatrix} 1.74716093 & 1.3103707 & 0.87358046 \\ 1.3103707 & 1.74716093 & 1.3103707 \\ 0.87358046 & 1.3103707 & 1.74716093 \end{pmatrix}. \tag{3.107}$$

For this choice of $\mathbf{C}$ $\rho_{1,2} = \rho_{2,3} = 0.75$, $\rho_{1,3} = 0.5$, and $|\mathbf{C}| = 1$. The covariance matrix of $\hat{S}$

is diagonal:

$$\hat{\mathbf{C}} = \begin{pmatrix} 0.283 & 0 & 0 \\ 0 & 0.868 & 0 \\ 0 & 0 & 4.032 \end{pmatrix}. \tag{3.108}$$

The receptive field configurations are the same as in Section 3.18.1. We note that in the $\vec{s}$ coordinates, all components have the same variance, whereas this symmetry is broken in the decorrelated components $\hat{s}$. We compared sorting the components of $\hat{s}$ in increasing and decreasing order of variance (triangles and squares respectively) in Figure 3.13A-C. Component order does not matter for $I_{\text{comp-ind}}(\hat{S}, \hat{T})$, Figure 3.13D. We find that for both I_{vector} and I_{iso} it is optimal to perform computation in the original basis, with both quantities accurately matching $\hat{I}(\vec{R}, \vec{S})$. For $I_{\text{comp-cond}}$ and $I_{\text{comp-ind}}$ accuracy is increased by using decorrelated components and, for $I_{\text{comp-cond}}$, sorting components by decreasing variance.

3.18.3 Highly asymmetric receptive field distributions

Next, we consider a small population ($N = 9$) with a highly asymmetric distribution of redundant receptive fields in two stimulus dimensions. In particular we are interested in a population where many different configurations of $\vec{R}$ map to the same configuration of $\vec{T}$, demonstrating the utility of using $\vec{T}$ as a non-trivial sufficient statistic of $\vec{R}$. With this in mind, we chose a heavily redundant configuration of $\{\vec{w}_n\}$: $\vec{w}_n = (0, 1)$ $(n = 1, 2, 3)$, $\vec{w}_n = (1, 0)$ $(n = 4, 5, 6)$, $\vec{w}_n = (1, 1)$ $(n = 7, 8, 9)$. The cardinalities of $\vec{R}$, $\vec{T}$, T_1 and T_2 are 512, 37, 7, and 7 respectively. Because N is small, ground truth values of $I(\vec{R}, \vec{S})$ were computed by exactly evaluating $P(\vec{r}|\vec{s})$ $\forall \vec{r} \in \{-1, 1\}^N$, for every sample of $\vec{s}$. Given $P(\vec{r}|\vec{s})$ we average across $\vec{s}$

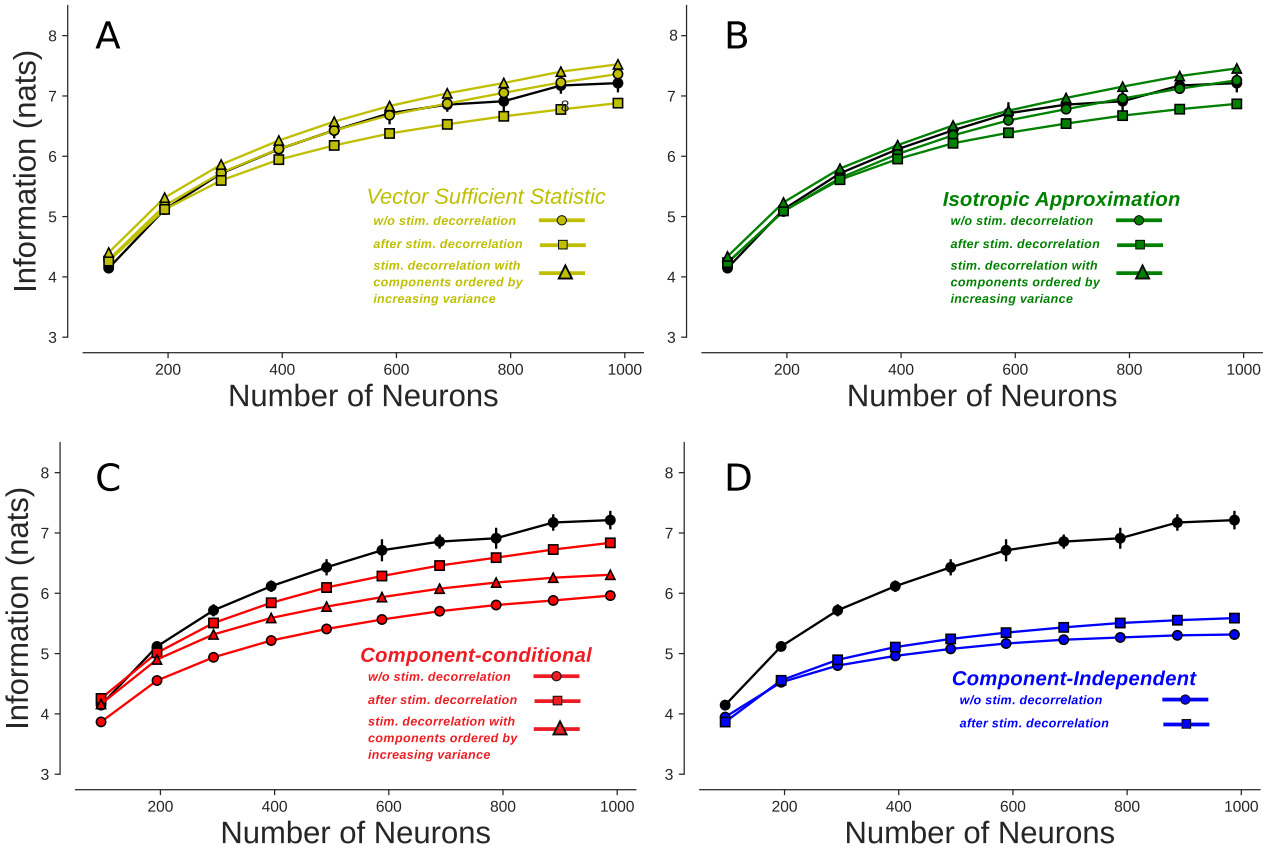

Figure 3.13: Information curves for neural populations with correlated RF distributions, cf. section 3.18.2. Lines and errorbars are mean and standard deviation over ten repeats of the estimator. In all panels, circles represent information computed in the original basis, while squares and triangles are computations performed in decorrelated basis. I_{vector} (**A**) and I_{iso} (**B**) recover full information. These two computations do not benefit from working in the decorrelated stimulus basis. Stimulus decorrelation improves the performanceof $I_{\text{comp-cond}}$ (**C**) and $I_{\text{comp-ind}}$ (**D**)). In (**D**), computations are invariant to ordering of components.

to get $P(\vec{r})$ explicitly, and calculate $H(\vec{R})$, $H(\vec{R}\,|\vec{S}\,)$, and $I(\vec{R},\vec{S}\,)$ from these distributions. $P(\vec{s})$ is gaussianly distributed with diagonal covariance matrix, in accordance with (3.100). We set $N_{\text{stim}} = 10,000$. To test how the relative values of $I(S_1, T_1)$ and $I(S_2, T_2)$ impact the optimal ordering of components in $I_{\text{comp-cond}}(\vec{S},\vec{T}\,)$ we fixed $\sigma_2 = 1$ and varied σ_1 between 0.5 and 2.5. Results are plotted in Figure 3.14. As predicted, the hierarchy of bounds (3.97) and (3.100) holds for all σ_1. It is also notable that in this case, just like in the case of large neural populations, for computing $I_{\text{comp-cond}}(\vec{S},\vec{T}\,)$, it always seems best to start the information computation with the stimulus component that has the largest variance. As expected, the ordering or components does not strongly impact $I_{\text{vector}}(\vec{S},\vec{T}\,)$.

3.18.4 Experimental stimuli and receptive fields

In the previous three experimental sections we considered synthetic distributions of low-dimensional stimuli and artificial configurations of receptive fields. In this section we analyze a population of model neurons with receptive fields and α values that were fit to responses of primary visual cortex neurons (V1) elicited by natural stimuli [Sharpee et al., 2006b]. We use 147 pairs of $(\vec{w}_n, \alpha_n)$ values that were fit using the Maximally Informative Dimension (MID) algorithm as in [Sharpee et al., 2006b]. Stimuli are 10 pixel by 10 pixel patches extracted from the same set of images used to fit the model parameters. Receptive fields are normalized and centered on the patch, and we chose a 10×10 sub-patch of the original 32×32 shaped receptive fields so that all dimensions are well sampled by receptive fields. That is, for all pixels of the 10×10 patch, at least 115 of the 147 neurons have a nonzero value in the corresponding component of their receptive field. Additionally, the

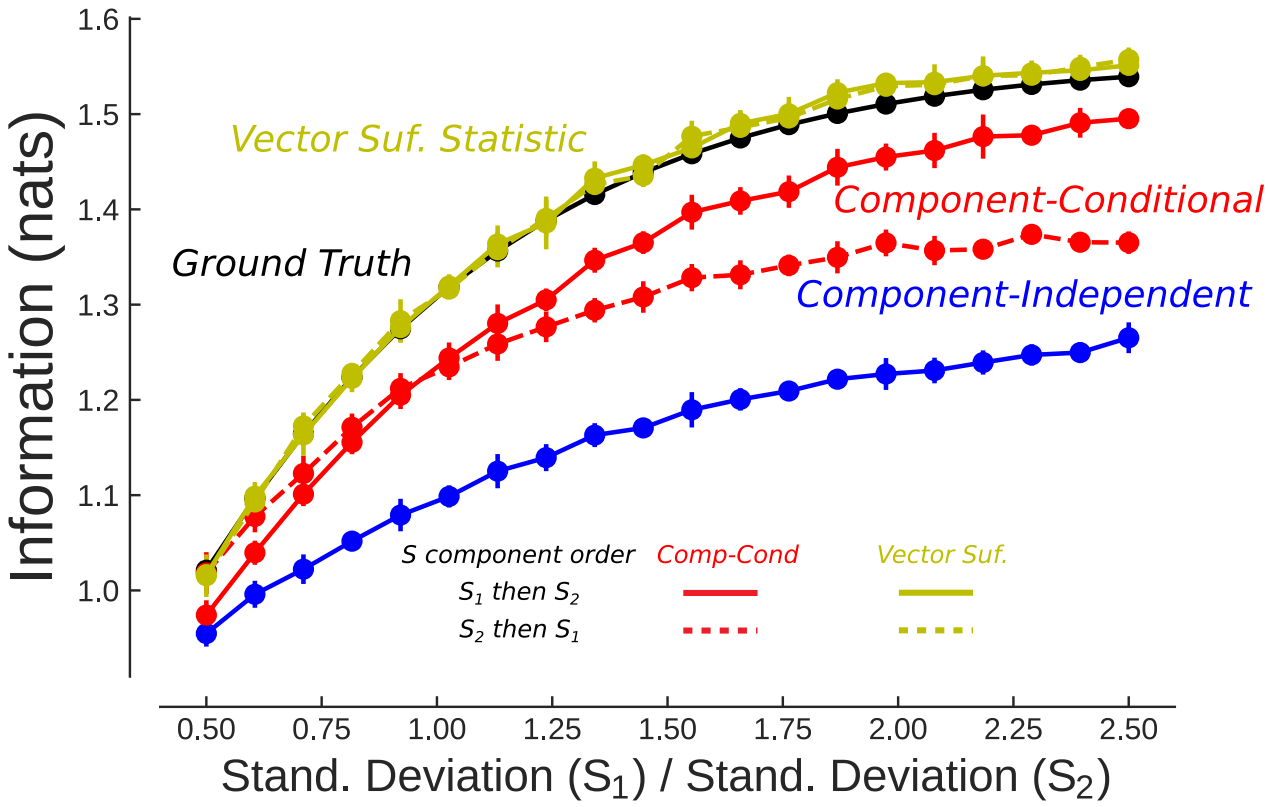

Figure 3.14: Information curves for the example population with highly redundant RFs from Sec. 3.18.3. Lines and errorbars are mean and standard deviation over ten repeats of the estimator. Although neither the component-conditional nor the component information are guaranteed to tightly approximate the information, both provide good approximation to the full information (as estimated via unbiased Monte Carlo method), reaching values within $\geq$ 80% of the maximum. For the vector-sufficient statistic, both component orderings accurately reproduced the full information. The next best approximation to the full information is provided by the component-conditional computation with components added in the order from largest to smallest variance. This approximation reaches accuracy within 95% of the full value over the range of neural population sizes.

stimuli are z-scored by subtracting the mean and dividing by the standard deviation, with both quantities computed across all samples and pixels collectively.

Because of high stimulus dimensionality we could only compute the $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ bound (via the KSG estimator) and the ground truth information (via the Monte Carlo method). Because the pixels of natural image patches are clearly not independent, we also computed $I_{\text{comp-ind}}$ in two additional coordinate systems. The first coordinate system is simply the linearly decorrelated components $\hat{S}$ and $\vec{T}$ defined in Eq. (3.105). The second coordinate systems uses independent components derived using independent component analysis on $\vec{S}$

$$\tilde{S} = \mathbf{U}\vec{S}, \quad , \tilde{T} = (\mathbf{U}^{-1})^T \vec{T} \tag{3.109}$$

Here, $\mathbf{U}$ is an unmixing matrix computed using Infomax ICA [Bell and Sejnowski, 1997] on the samples of $\vec{S}$. As is done in [Bell and Sejnowski, 1997], $\mathbf{U}$ includes a linear whitening matrix. We note that in Eq. (3.109), $\vec{T}$ is multiplied by $(\mathbf{U}^{-1})^T$ and not $\mathbf{U}$ because the ICA unmixing matrix is generally not orthogonal and we require that $\vec{t} \cdot \vec{s} = \tilde{t} \cdot \tilde{s}$. As before $I_{\text{vector}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\hat{S}, \hat{T}) = I_{\text{vector}}(\tilde{S}, \tilde{T})$. However, the same cannot be said for $I_{\text{comp-ind}}$ expressed in different coordinate systems.

To evaluate the effect of using different coordinate systems to evaluate $I_{\text{comp-ind}}$ for different population sizes we first ranked the 147 neurons in descending order by the information each neuron carried about the stimulus. We computed $I(R_n, \vec{S}) \; \forall n \in \{1, ..., 147\}$, which is easily done exactly since R_n is a binary variable, and then sorted neurons so that $I(R_n, \vec{S}) >= I(R_{n+1}, \vec{S}) \; \forall n$. We considered populations of size $N = 60, 70, ..., 140$, where each population contained the first N neurons under the aforementioned ordering. For each

value of N we computed $\hat{I}(\vec{R}, \vec{S}\,)$ $(B = 3)$, $I_{\text{comp-ind}}(\vec{S}, \vec{T}\,)$, $I_{\text{comp-ind}}(\hat{S}, \hat{T})$, and $I_{\text{comp-ind}}(\tilde{S}, \tilde{T})$. We note that $\log(N_{\text{stim}}) = \log(49, 152) \approx 10.8$ nats. Results are plotted in Figure 3.15.

We observe that both $I_{\text{comp-ind}}(\vec{S}, \vec{T}\,)$ and $I_{\text{comp-ind}}(\hat{S}, \hat{T})$ overestimate the true information, especially $I_{\text{comp-ind}}(\vec{S}, \vec{T}\,)$. This overestimation occurs because stimulus components are not independent. By comparison, computation performed in the ICA basis, $I_{\text{comp-ind}}(\tilde{S}, \tilde{T})$, lower bounds the mutual information for all N, achieving $\geq 75\%$ of the full information across the range of population sizes.

3.18.5 Handling intrinsically correlated neurons

In order to simplify derivations, we assumed that the neural responses were independent after conditioning on $\vec{s}$. However, all of the analytic results in Section 3.17.2 can be extended to specific forms of intrinsic interneuronal correlation to allow for the presence of correlations in neural responses for a given stimulus $\vec{s}$. Formally, we modify the base measure $h(\vec{r}\,)$ to include a pairwise coupling term:

$$h(\vec{r}, \mathbf{J}) = e^{\sum_{mn} \mathbf{J}_{mn} r_m r_n - \sum_n r_n \alpha_n}. \tag{3.110}$$

In Eq. 3.110, $\mathbf{J}$ is a symmetric $N \times N$ matrix where J_{mn} describes the intrinsic coupling between the m^{th} and n^{th} neurons. In this case $P(\vec{r}\,|\vec{s}, \mathbf{J})$ can still be written as an exponential

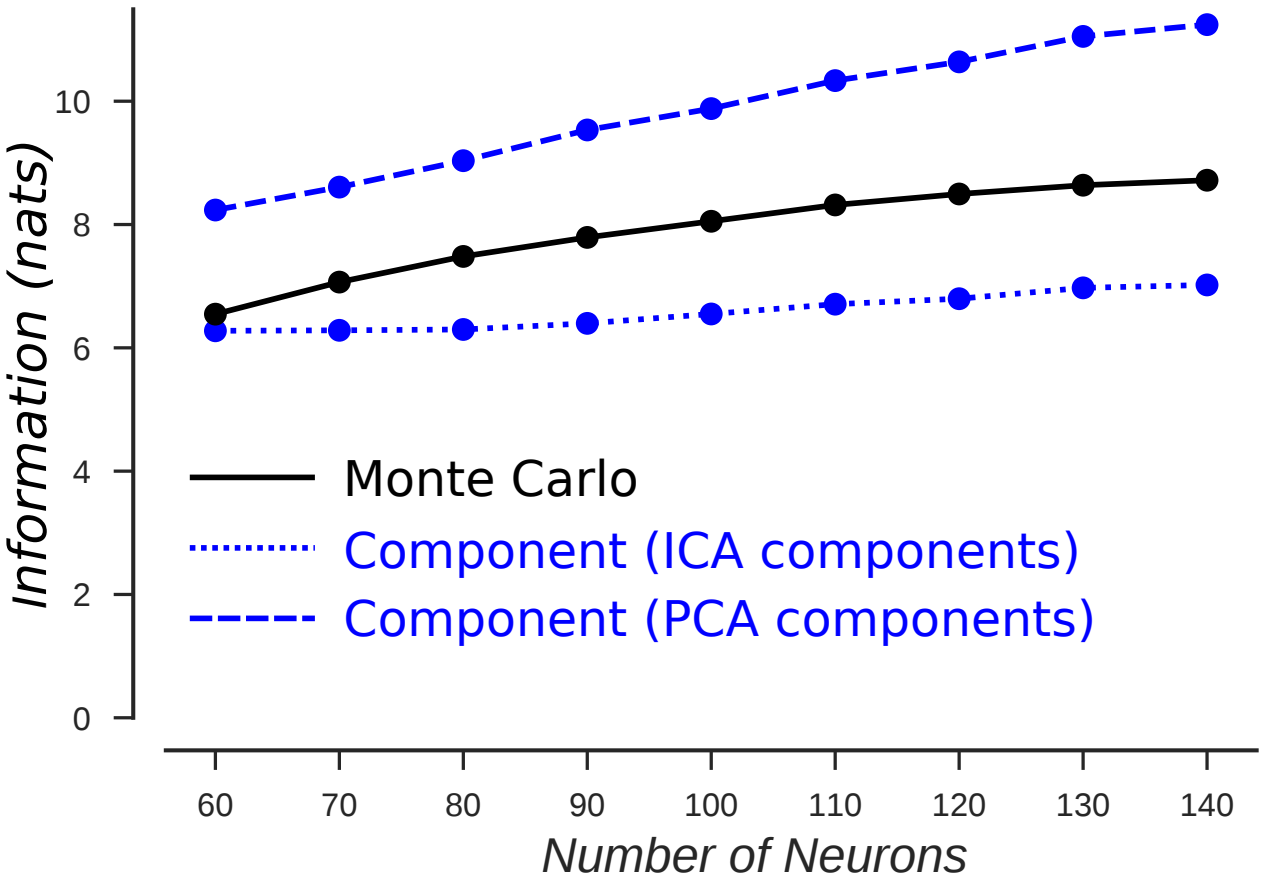

Figure 3.15: Information curves for populations based on experimentally recorded RFs and probed with $D = 100$ natural visual stimuli. The full information (solid black line) is computed via the Monte Carlo method and is compared to $I_{\text{comp-ind}}$ approximations computed in two different bases: the PCA basis (blue dashed line) and the ICA basis (blue dotted line). We do not show the calculation in the original pixel basis, because it$I_{\text{comp-ind}}(\vec{S}, \vec{T})$ is omitted as it was yielded values $\sim$ 25 nats across the range of population sizes and obscured the other curves. Because of non-Gaussian statistics of natural stimuli, PCA components remain correlated, and as a result the approximation is no longer guaranteed to be a lower bound of the true information. Computation in the ICA basis respects the lower bound requirements, and achieves $\geq 75\%$ of the full information across the range of population sizes.

family in a canonical form:

$$\begin{aligned} P(\vec{r}|\vec{s}) &= h(\vec{r},\mathbf{J})\exp(\vec{s}\cdot\vec{t}(\vec{r})-A(\vec{s},\mathbf{J})), \\ A(\vec{s},\mathbf{J}) &= \log\left(\sum_{\vec{r}} h(\vec{r},\mathbf{J})\exp(\vec{s}\cdot\vec{t}(\vec{r}))\right). \end{aligned} \tag{3.111}$$

The form of the sufficient statistic remains unchanged, though $A(\vec{s},\mathbf{J})$ generally lacks a closed form. Nevertheless, all of the decompositions, equalities, and inequalities in Section 3.17.2 require only for the exponential family to be in canonical form and remain valid.

We tested the accuracy of our approximations on a small ($N = 10$) population of intrinsically correlated neurons. Receptive fields are uniformly distributed on the unit circle, $\alpha_n = 0\ \forall\, n$, and $P(\vec{s})$ is a standard two-dimensional Gaussian ($N_{\text{stim}} = 20,000$). Intrinsic coupling is set proportional to the overlap between receptive fields with a coupling strength J_0, the sign of which determines whether the intrinsic couplings perform stimulus decorrelation or error-correction [Tkačik et al., 2010].:

$$\mathbf{J}_{mn} = J_0\, \vec{w}_m \cdot \vec{w}_n. \tag{3.112}$$

The algorithms of Sections 3.18 and 3.16 all depend on being able to sample easily from $P(\vec{r}|\vec{s})$. For large N and general $\mathbf{J}$ this is usually difficult, particularly for configurations of $\mathbf{J}$ that exhibit glassy dynamics. Additionally, evaluating Eq. (3.77) requires explicit knowledge of $A(\vec{s})$, though methods such as mean-field theory or the TAP approximation may be used to approximate $A(\vec{s})$ [Opper et al., 2001]. Since this population is small, we evaluate the ground truth information exactly as in Section 3.18.3. Likewise, we sample $\vec{r}$ from $P(\vec{r}|\vec{s})$ exactly by

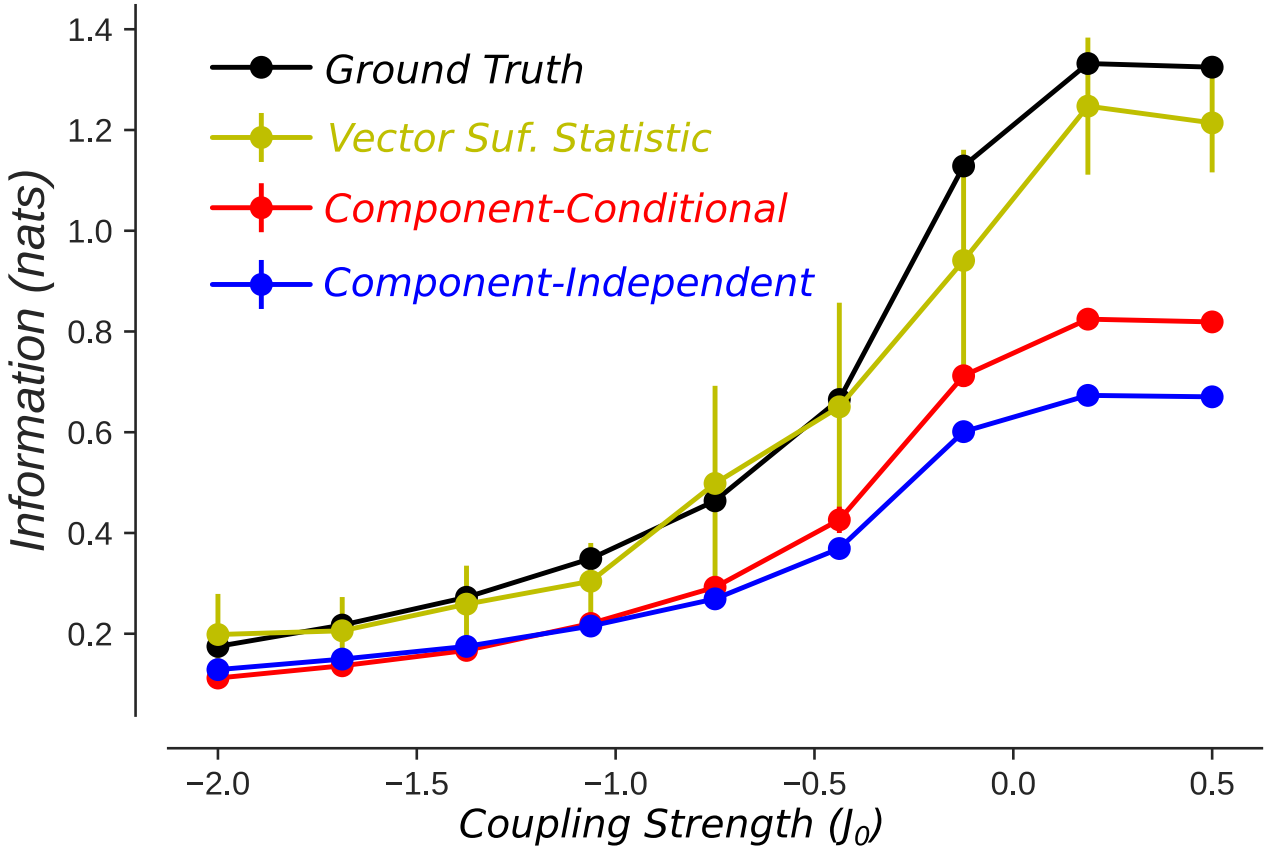

Figure 3.16: Information curves for populations with intrinsic correlations. Lines and errorbars are mean and standard deviation over ten repeats of the estimator.

computing all $2^N\,(1,024)$ probabilities for every sample of $\vec{s}$. Analyzing the error introduced in using approximate sampling strategies such as Markov Chain Monte Carlo is left to future investigation. Results are plotted in Figure 3.16. As predicted, $I_{\text{vector}}(\vec{S},\vec{T}\,)$ matches the ground truth values of the information. Similarly the hierarchy of bounds (3.97) and (3.100) is preserved, though for strongly negative couplings $I_{\text{comp-cond}}(\vec{S},\vec{T}\,) \approx I_{\text{comp-ind}}(\vec{S},\vec{T}\,)$. In sum, the presence of noise correlations does not invalidate the approximations and bounds that are derived above. However, numerical computation can become more difficult in the presence of noise correlations.

3.19 Conclusions and Future Work

We have presented three approximations that can be used to estimate the information transmitted by large neural populations. Each of these three approximations represents different trade-offs between accuracy and computational ease and feasibility. The best performance in terms of accuracy was provided by the isotropic approximation I_{iso}. This approximation worked well even in cases where is not guaranteed to become asymptotically exact with increasing population size. For example, the isotropic approximation was derived assuming a matching covariance matrix for both the stimulus and RF distributions, cf. Appendix 3.21. Yet, this approximation matched the full ground-truth information values even for correlated stimuli with RFs remaining uncorrelated across the neural population (Fig. 3.13.) This approximation provided the best overall performance among the bounds tested, consistent with the theoretically expected inequalities between these bounds, cf. Eq. (3.97).

The component-conditional information $I_{\text{comp-cond}}$ offered the second-best performance. This approximation performed especially well when computed in the stimulus basis where stimulus components were not correlated. This approximation is less computationally difficult compared to I_{iso}, because each conditional information is evaluated between just two quantities S_d and T_d compared to S_d and a conjunction of T_d and $|T_{>d}|$ as in I_{iso}. For this reason, the finite-sample bias of $I_{\text{comp-ind}}$ can also be less than I_{iso}, because bias in the evaluation of the mutual information is usually larger for higher-dimensional calculations.

The last approximation $I_{\text{comp-ind}}$ is the least accurate of the three approximation but is computationally the easiest. It is the only approximation among the three we considered

here that we were able to implement in conjunction with high dimensional stimuli. This approximation becomes most accurate in the stimulus bases where stimulus components are independent. There is strong evidence that neural receptive fields are organized along the ICA components of natural stimuli [Bell and Sejnowski, 1997, Olshausen and Field, 2004, Smith and Lewicki, 2006]. This raises the possibility that the approaches proposed here will fair well when applied to recorded neural responses. Indeed, we found that $I_{\text{comp-ind}} \geq 75\%$ of the full information value for large neural populations constructed using experimentally recorded RFs and probed with natural stimuli.

At present, the main limitation for computing the conditional approximations $I_{\text{comp-cond}}$ and $I_{\text{comp-iso}}$ is not the number of neurons but rather the stimulus dimensionality. For stimulus distributions where $P(s_{\geq d}|s_{<d})$ can be easily sampled from, such as Gaussian distributions, we can take advantage of the fourth line of Eq. (3.88) to compute unbiased estimates of $I_{\text{comp-cond}}$ and I_{iso}, albeit with possibly high variance. Developing methods that can efficiently approximate these conditional computations represents an important opportunity for future research.

3.20 Appendix A: Bias of $\hat{H}(\vec{R})$

In this section we give a self-contained proof that $\hat{H}(\vec{R}\,)$ systematically underestimates the "true" entropy $H(\vec{R}\,)$. We first assume that $P(\{\vec{s}_\mu\}) = \prod_\mu P(\vec{s}_\mu)$: The $\vec{s}_\mu$ are drawn independently from $P(\vec{s}\,)$, whether $P(\vec{s}\,)$ is a smooth density on $\mathcal{R}^D$ or some larger set of samples. We define a an empirical version of the marginal distribution on $P(\vec{r}\,)$:

$$\hat{P}(\vec{r}|\{\vec{s}_\mu\}) = \frac{1}{M}\sum_\mu P(\vec{r}|\vec{s}_\mu) \tag{3.113}$$

The true marginal distribution is the expected value of $\hat{P}(\vec{r}|\{\vec{s}_\mu\})$: $\langle \hat{P}(\vec{r};\{\vec{s}_\mu\})\rangle_{P(\{\vec{s}_\mu\})} = P(\vec{r})$. The shannon entropy is a concave function of $\hat{P}(\vec{r};\{\vec{s}_\mu\})$, which can be considered a random vector in the 2^N dimensional probability simplex. Thus, by Jensen's inequality we have the following:

$$\left\langle \hat{H}(\vec{R}) \right\rangle_{P(\{\vec{s}_\mu\})} \leq H(\vec{R} : P = \langle \hat{P} \rangle) \equiv H(\vec{R}) \tag{3.114}$$

This bias holds even in the case of evaluating $\hat{H}(\vec{R})$ through exact enumeration. We note that we are able to produce unbiased estimates of $\hat{H}(\vec{R})$ because we have full access to $P(\vec{r}|\vec{s}_\mu)$: We can evaluate $P(\vec{r}|\vec{s}_\mu)$ explicitly and deterministically, and thus $\hat{F}(\vec{r})$ as well (up to factors of numerical precision). If we were always constrained to drawing samples from $P(\vec{r}|\vec{s}_\mu)$, then we would once again be limited to making biased estimates of $\hat{H}(\vec{R})$ [Paninski, 2003].

3.21 Appendix B: On the asymptotic tightness of $I_{\mathbf{iso}}(\vec{S}, \vec{T})$

Consider a large population ($N \gg 1$) where the distribution of $\vec{w}$ and α is such that $A(\vec{s}) = A(|\vec{s}|)$ (in some sense to be made more precise later). Consider the likelihood ratio in the definition of (3.93):

$$\begin{aligned} \frac{P(t_{\geq d}|s_d, s_{<d})}{P(t_{\geq d}|s_{<d})} &= \frac{\langle \exp(s_d t_d + s_{>d}\cdot t_{>d} - A(\vec{s}))\rangle_{s_{>d}|s_{\leq d}}}{\langle \exp(s_d t_d + s_{>d}\cdot t_{>d} - A(\vec{s}))\rangle_{s_{\geq d}|s_{<d}}} \\ &= \frac{\langle \exp(s_d t_d + s_{>d}\cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{>d}|s_{\leq d}}}{\langle \exp(s_d t_d + s_{>d}\cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{\geq d}|s_{<d}}} \end{aligned} \tag{3.115}$$

Additionally we consider a stimulus distribution that is similarly isotropic so that the conditional distribution $P(s_{>d}|s_{\leq d})$ can be written in a convenient factored form:

$$\begin{aligned} P(\vec{s}) &= P(|\vec{s}|) = P\left(\sqrt{s_{<d}^2 + s_d^2 + s_{>d}^2}\right) \\ P(s_{>d}|s_{\leq d}) &= \frac{P(\vec{s})}{P(s_{\leq d})} = \frac{P\left(\sqrt{s_{<d}^2 + s_d^2 + s_{>d}^2}\right)}{P(s_{\leq d})} \end{aligned}$$

We will show that in this situation, we can replace $T_{\geq d}$ in (3.93) with the variable that is the concatenation of T_d and $|T_{>d}|$ without loss of information:

$$I(S_d, T_{\geq d}|s_{<d}) \approx I(S_d, \{T_d, |T_{>d}|\}\,|s_{<d}) \tag{3.116}$$

To show this using the Fisher-Neyman factorization theorem, it suffices to show that the numerator and denominator in (3.115) can be factored as follows:

$$\begin{aligned} \langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{>d}|s_{\leq d}} &= g_1(s_{<d}, s_d, t_d, |t_{>d}|) g_2(t_{>d}) \\ \langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{\geq d}|s_{<d}} &= f_1(s_{<d}, t_d, |t_{>d}|) f_2(t_{>d}) \end{aligned} \tag{3.117}$$

With the requirement that $f_2(t_{>d}) = g_2(t_{>d})$, so that dependence on $t_{>d}$ cancels out in (3.115). We note that the first line of (3.117) implies the second so we examine that term in more detail.

$$\begin{aligned} \langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{>d}|s_{\leq d}} &= \exp(s_d t_d) \int \exp\left(s_{>d} \cdot t_{>d} - A\left(\sqrt{s_{<d}^2 + s_d^2 + s_{>d}^2}\right)\right) \\ &\times \; P(s_{>d}|s_{\leq d}) ds_{>d} \\ &= \frac{\exp(s_d t_d)}{P(s_{\leq d})} \int \exp\left(s_{>d} \cdot t_{>d} - A\left(\sqrt{s_{<d}^2 + s_d^2 + s_{>d}^2}\right)\right) \\ &\times \; P\left(\sqrt{s_{<d}^2 + s_d^2 + s_{>d}^2}\right) ds_{>d} \end{aligned} \tag{3.118}$$

We note that $s_{>d}$ is a $K = D - d$ dimensional vector. We assume that $d < D - 1$, so that $K > 1$, otherwise no further reduction of (3.118) is possible. We convert the integral over $\mathcal{R}^K$ in (3.118) into spherical coordinates and break it into three parts: Integration over $\rho \in [0, \infty)$ where $|s_{>d}| = \rho$; integration over $\theta \in [0, \pi]$ where $s_{>d} \cdot t_{>d} = \rho |t_{>d}| \cos(\theta)$, and $\varphi \in \Omega_{K-1}$ is the set of all directions in $\mathcal{R}^K$ with constant θ. The integrand of (3.118) doesn't depend on φ so we can integrate over it automatically, yielding a constant B_K that is a function only of K. We can now restate (3.118) in these coordinates:

$$\begin{aligned}
\ldots &= \frac{B_K \exp(s_d t_d)}{P(s_{\leq d})} \int_0^\infty \int_0^\pi d\rho d\theta \rho^{K-1} \sin^{K-2}(\theta) P\left(\sqrt{s_{<d}^2 + s_d^2 + \rho^2}\right) \\
&\times \exp\left(\rho |t_{>d}| \cos(\theta) - A\left(\sqrt{s_{<d}^2 + s_d^2 + \rho^2}\right)\right) \\
&= \frac{B_K \exp(s_d t_d)}{P(s_{\leq d})} \int_0^\infty d\rho \rho^{K-1} P\left(\sqrt{s_{<d}^2 + s_d^2 + \rho^2}\right) \\
&\times \exp\left(-A\left(\sqrt{s_{<d}^2 + s_d^2 + \rho^2}\right)\right) \int_0^\pi d\theta \sin^{K-2}(\theta) \exp\left(\rho |t_{>d}| \cos(\theta)\right) \qquad (3.119)
\end{aligned}$$

We next evaluate the integral over θ in (3.119).

$$\int_0^\pi d\theta \sin^{K-2}(\theta) \exp(\rho |t_{>d}| \cos(\theta)) = \sqrt{\pi} \frac{\Gamma\left(\frac{K}{2} - \frac{1}{2}\right)}{\Gamma\left(\frac{K}{2}\right)} {}_0F_1\left(\frac{K}{2}, \frac{\rho^2 |t_{>d}|^2}{4}\right) \equiv F_K(\rho |t_{>d}|) \qquad (3.120)$$

Where $\Gamma(x)$ is the Gamma function, and ${}_0F_1(a, z)$ is the confluent hypergeometric limit function. We have our final expression for the first term in (3.117):

$$\begin{aligned}
\ldots &= \frac{B_K \exp(s_d t_d)}{P(s_{\leq d})} \int_0^\infty d\rho \rho^{K-1} P\left(\sqrt{s_{<d}^2 + s_d^2 + \rho^2}\right) \exp\left(-A\left(\sqrt{s_{<d}^2 + s_d^2 + \rho^2}\right)\right) F_K(\rho |t_{>d}|) \\
&= \frac{B_K \exp(s_d t_d)}{P(s_{\leq d})} g(s_{<d}, s_d, |t_{>d}|) \qquad (3.121)
\end{aligned}$$

By setting g_1 in (3.117) equal (3.121), and letting $g_2 = f_2 = 1$, we have established (3.116).

3.21.1 Approximating $A(\vec{s})$ for Gaussian $P(\vec{w})$

In the previous section we assumed that the distribution $P(\vec{w})$ and $P(\alpha)$ are such that $A(\vec{s}) = A(|\vec{s}|)$. In the special case when $N \gg 1$, $P(\alpha) = \delta(\alpha)$, and $\vec{w}$ are Gaussianly distributed we can approximate $A(\vec{s})$ in a semi-closed form. Let $P(\vec{w})$ be a zero-mean Gaussian with with positive-definite covariance matrix C:

$$\begin{aligned} A(\vec{s}) &= N \int d\vec{w} \frac{\exp\left(-\frac{1}{2}\vec{w}^T C^{-1} \vec{w}\right)}{\sqrt{\det(2\pi C)}} \log(2\cosh(\vec{w}\cdot\vec{s})) \\ &= N \int dx \frac{\exp\left(\frac{-x^2}{2\sigma_x^2}\right)}{\sqrt{2\pi\sigma_x^2}} \log(2\cosh(x)) \qquad (3.122) \\ \sigma_x &= \sqrt{\vec{s}^T C \vec{s}} \qquad (3.123) \end{aligned}$$

Where we have taken advantage of the fact that $\vec{w}\cdot\vec{s}$ is a scalar gaussian variable with standard deviation that depends on $\vec{s}$ and C. We next take an infinite series expansion of $\log(2\cosh(x))$.

$$\log(2\cosh(x)) = |x| + \log\left(1 + \exp(-2|x|)\right) = |x| + \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \exp(-2m|x|) \qquad (3.124)$$

As an aside, the first equality in (3.124) is a useful and numerically stable expression for $A(x)$. The "softplus" function $l(y) = \log(1 + \exp(y))$ is implemented in many scientific computing packages, and using this alternate form for $A(x)$ sidesteps computing the hyperbolic cosine. We next take the appropriate Gaussian average of each term in (3.124):

$$\begin{aligned} \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{\infty} dx \exp\left(\frac{-x^2}{2\sigma_x^2}\right) |x| &= \sqrt{\frac{2}{\pi}}\sigma_x \qquad (3.125) \\ \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{\infty} dx \exp\left(\frac{-x^2}{2\sigma_x^2} - 2m|x|\right) &= \mathrm{erfcx}(\sqrt{2}m\sigma_x) \qquad (3.126) \end{aligned}$$

Where erfcx(y) is the scaled complementary error function. Thus we have our final form for $A(\vec{s})$:

$$A(\vec{s}) = N\sqrt{\frac{2}{\pi}}\sigma_x + N\sum_{m=1}^{\infty}\frac{(-1)^{m+1}}{m}\text{erfcx}(\sqrt{2}m\sigma_x) \tag{3.127}$$

We note that erfcx(y) monotonically decreases to zero, so for large values of $\sqrt{\vec{s}^T C\vec{s}}$, $A(\vec{s})$ is well approximated by the first term in (3.127). Regardless we see that $A(\vec{s})$ depends on $\vec{s}$ only through $\sqrt{\vec{s}^T C\vec{s}}$:

$$A(\vec{s}) = A\left(\sqrt{\vec{s}^T C\vec{s}}\right) = A\left(|\mathbf{U}\vec{s}|\right) \tag{3.128}$$

Where $\mathbf{U} = C^{\frac{1}{2}}$ is the cholesky decomposition of C.

3.21.2 Generalizing $I_{\text{iso}}(\vec{S},\vec{T}\,)$ for matched anisotropy

In this section we will show that a generalized form of $I_{\text{iso}}(\vec{S},\vec{T}\,)$ will also asymptotically converge to $I_{\text{vector}}(\vec{S},\vec{T}\,)$ when both $P(\vec{s})$ and $A(\vec{s})$ obey a certain form of "matched" anisotropy. Specifically we assume that $P(\vec{s})$ and $A(\vec{s})$ depend on on $\vec{s}$ through a quadratic function of $\vec{s}$ with positive-definite kernel $\mathbf{C}$.

$$P(\vec{s}) = P\left(\sqrt{\vec{s}^T C\vec{s}}\right) = P\left(|\mathbf{U}\vec{s}|\right) \tag{3.129}$$

$$A(\vec{s}) = A\left(\sqrt{\vec{s}^T C\vec{s}}\right) = A\left(|\mathbf{U}\vec{s}|\right) \tag{3.130}$$

Where, as in Section 3.21.1, $\mathbf{U} = C^{\frac{1}{2}}$ is the cholesky decomposition of C. Let us transformed versions of $\vec{S}$ and $\vec{T}$:

$$\tilde{S} = \mathbf{U}\vec{S} \tag{3.131}$$

$$\tilde{T} = \mathbf{U}^{-T}\vec{T} \tag{3.132}$$

Where $\mathbf{U}^{-T}$ is the transpose of the inverse of $\mathbf{U}$, which is well defined since C was positive definite. We note that $\tilde{t} \cdot \tilde{s} = \vec{t} \cdot \vec{s}$ and $P(\tilde{t}|\tilde{s})$ can once again be written as an exponential family in canonical form:

$$P(\tilde{t}|\tilde{s}) = \exp(\tilde{s} \cdot \tilde{t} - A(\tilde{s}))h(\tilde{t}) \tag{3.133}$$

As the mappings from $\vec{S}$ to $\tilde{S}$ and $\vec{T}$ to $\tilde{T}$ are diffeomorphisms, we have that $I_{\text{vector}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\tilde{S}, \tilde{T})$. Furthermore, equation (3.133) implies that an analogous form of equation (3.94) holds for $I_{\text{vector}}(\tilde{S}, \tilde{T})$:

$$I(\tilde{S}_d, \tilde{T}|\tilde{S}_{<d}) = I(\tilde{S}_d, \tilde{T}_{\geq d}|\tilde{S}_{<d}) \tag{3.134}$$

Additionally, $A(\tilde{s}) = A(|\tilde{s}|)$ and $P(\tilde{s}) = P(|\tilde{s}|)$. Thus, we may reuse the derivation of section 3.116 to derive the following analogy of equation (3.116):

$$I(\tilde{S}_d, \tilde{T}_{\geq d}|\tilde{s}_{<d}) \approx I\left(\tilde{S}_d, \left\{\tilde{T}_d, |\tilde{T}_{>d}|\right\} \middle| \tilde{s}_{<d}\right) \tag{3.135}$$

Therefore $I_{\text{iso}}(\tilde{S}, \tilde{T})$ is asymptotically equal to $I_{\text{vector}}(\vec{S}, \vec{T})$:

$$I_{\text{iso}}(\tilde{S}, \tilde{T}) = \sum_d I\left(\tilde{S}_d, \left\{\tilde{T}_d, |\tilde{T}_{>d}|\right\} \middle| \tilde{S}_{<d}\right) \approx I_{\text{vector}}(\vec{S}, \vec{T}) \tag{3.136}$$

We note that an example of such a matched isotropy situation would be where both the stimuli and receptive fields (for a large population) are distributed according to a gaussian distribution with covariance matrix C (c.f. section 3.21.1).

3.22 Appendix C: Independent Sub-populations

In this section we present an example where one of our proposed approximations, $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ in this case, is equal to $I_{\text{vector}}(\vec{S}, \vec{T})$. Let $(\hat{e}'_1, ..., \hat{e}'_D)$ be an orthonormal basis for $\mathcal{R}^D$. Suppose that the distribution of $\vec{s}$ and $\vec{w}$ are such that both $P(\vec{s})$ and $A(\vec{s})$ factor when expressed in this basis $(s'_k = \vec{s} \cdot \hat{e}'_k)$:

Similarly defining $t'_k = \vec{t} \cdot \hat{e}'_k$ we have that $\vec{t} \cdot \vec{s} = \vec{t}' \cdot \vec{s}'$. Because mutual information is invariant under bijective transformations of the variables (e.g. a change of basis) [Cover and Thomas, 2012] we have that $I_{\text{vector}}(\vec{S}, \vec{T}) = I(\vec{S}', \vec{T}')$. It is easy to show that $P(\vec{t}'|\vec{s}')$ can be written as follows:

$$P(\vec{t}'|\vec{s}') = h(\vec{t}') \prod_k \exp(s'_k t'_k - A(s'_k)) \tag{3.137}$$

Eq (3.137) implies that the log-likelihood ratio of $P(\vec{t}'|\vec{s}')$ to $P(\vec{t}')$ decomposes across s'_k:

$$\log\left(\frac{P(\vec{t}'|\vec{s}')}{P(\vec{t}')}\right) = \sum_k \log\left(\frac{P(t'_k|s'_k)}{P(t'_k)}\right) \tag{3.138}$$

Thus we have the following reduction of $I_{\text{vector}}(\vec{S}, \vec{T})$:

$$I_{\text{vector}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\vec{S}', \vec{T}') = \sum_k I(S'_k, T'_k) = I_{\text{comp-ind}}(\vec{S}', \vec{T}') \tag{3.139}$$

We note that (3.99) includes the case where for some k, $\vec{w}_n \cdot \hat{e}'_k = 0 \, \forall n$. In such a case $A(s'_k) = A(0) = \log(2)$, $t'_k = 0$ with probability one, and $I(S'_k, T'_k) = 0$. Thus, in the case of independent subpopulations, $I_{\text{vector}}(\vec{S}, \vec{T})$ can be reduced to computing $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ following a change of basis.

3.23 Appendix D: Relationship between $I_{\text{k-w}}(\vec{R}, \vec{S}\,)$ and $I_{\text{Fisher}}(\vec{R}, \vec{S}\,)$

In this appendix we relate $I_{\text{k-w}}(\vec{R}, \vec{S}\,)$ to the Fisher Information based approximation of [Brunel and Nadal, 1998]:

$$I_{\text{Fisher}}(\vec{R}, \vec{S}\,) = H(\vec{S}\,) + \frac{1}{2}\int d\vec{s} P(\vec{s}\,) \log\left(\frac{|\mathbf{J}(\vec{s}\,)|}{(2\pi e)^D}\right) \tag{3.140}$$

Where $\mathbf{J}(\vec{s}\,)$ is the Fisher Information Matrix of $P(\vec{r}\,|\vec{s}\,)$:

$$\mathbf{J}_{ab}(\vec{s}\,) = \left\langle \frac{\partial^2}{\partial_a \partial_b} \log P(\vec{r}\,|\vec{s}\,) \right\rangle_{P(\vec{r}\,|\vec{s}\,)} \tag{3.141}$$

We begin by considering the inner expectation over $\vec{s}\,'$ in $I_{\text{k-w}}(\vec{R}, \vec{S}\,)$:

$$L(\vec{s}\,) = \int d\vec{s}\,' P(\vec{s}\,') \exp\left(-D_{KL}(P(\vec{r}\,|\vec{s}\,)||P(\vec{r}\,|\vec{s}\,'))\right) \tag{3.142}$$

We next assume that the activation function $f_n(\vec{s}\,)$ is affine (e.g. (3.81)), and thus in canonical form. We also assume that, $P(\vec{r}\,|\vec{s}\,)$ is identifiable:

$$D_{KL}(P(\vec{r}\,|\vec{s}\,)||P(\vec{r}\,|\vec{s}\,') = 0 \Leftrightarrow \vec{s} = \vec{s}\,' \tag{3.143}$$

For (3.81) a necessary and sufficient condition for identifiability is that the matrix $\mathbf{W}$ has full rank, a reasonable assumption when $N \gg D$. We utilize the following properties of exponential families in canonical form:

1. $\frac{\partial^2}{\partial'_a \partial'_b} D_{KL}(P(\vec{r}|\vec{s})||P(\vec{r}|\vec{s}') = \frac{\partial^2}{\partial'_a \partial'_b} A(\vec{s}') = \mathbf{J}(\vec{s}')$.

2. $A(\vec{s}')$ is convex and $\mathbf{J}(\vec{s}')$ is positive semi-definite. When $P(\vec{r}|\vec{s})$ is identifiable replace $A(\vec{s}')$ becomes strictly convex, $\mathbf{J}(\vec{s}')$ positive definite, and $D_{KL}(P(\vec{r}|\vec{s})||P(\vec{r}|\vec{s}')$ has a global minimum with respect to $\vec{s}'$ of 0 at $\vec{s}' = \vec{s}$.

In the limit $N \gg D$ we approximate $L(\vec{s})$ using Laplace's Method [Bender and Orszag, 1999], expanding around $\vec{s}' = \vec{s}$:

$$L(\vec{s}) \approx P(\vec{s}) \sqrt{\frac{(2\pi)^D}{|\mathbf{J}(\vec{s})|}} \tag{3.144}$$

Plugging (3.144) into the definition of $I_{\text{k-w}}(\vec{R}, \vec{S})$ we have the following asymptotic expression for $I_{\text{k-w}}(\vec{R}, \vec{S})$:

$$\begin{aligned} I_{\text{k-w}}(\vec{R}, \vec{S}) &\approx H(\vec{s}) + \frac{1}{2} \int d\vec{s} P(\vec{s}) \log \left(\frac{|\mathbf{J}(\vec{s})|}{(2\pi)^D} \right) \\ &= I_{\text{Fisher}}(\vec{R}, \vec{S}) - \frac{D}{2} \end{aligned} \tag{3.145}$$

For stimulus distributions where the entropy $H(\vec{S})$ is known *a priori*, such as the gaussian distributions in sections 3.18.1 and 3.18.2, $I_{\text{Fisher}}(\vec{R}, \vec{S})$ can be computed in $O(M)$ time. If not, then $H(\vec{S})$ must be estimated, a challenging task in high dimensions. In figure 3.17, we replot the results of sections 3.18.1 with the inclusion of $I_{\text{Fisher}}(\vec{R}, \vec{S})$. We see that $I_{\text{Fisher}}(\vec{R}, \vec{S})$ is a very loose upper bound of $I(\vec{R}, \vec{S})$ and of $I_{\text{k-w}}(\vec{R}, \vec{S})$, indicating that the convergence of Laplace's Method may be very slow in this situation.

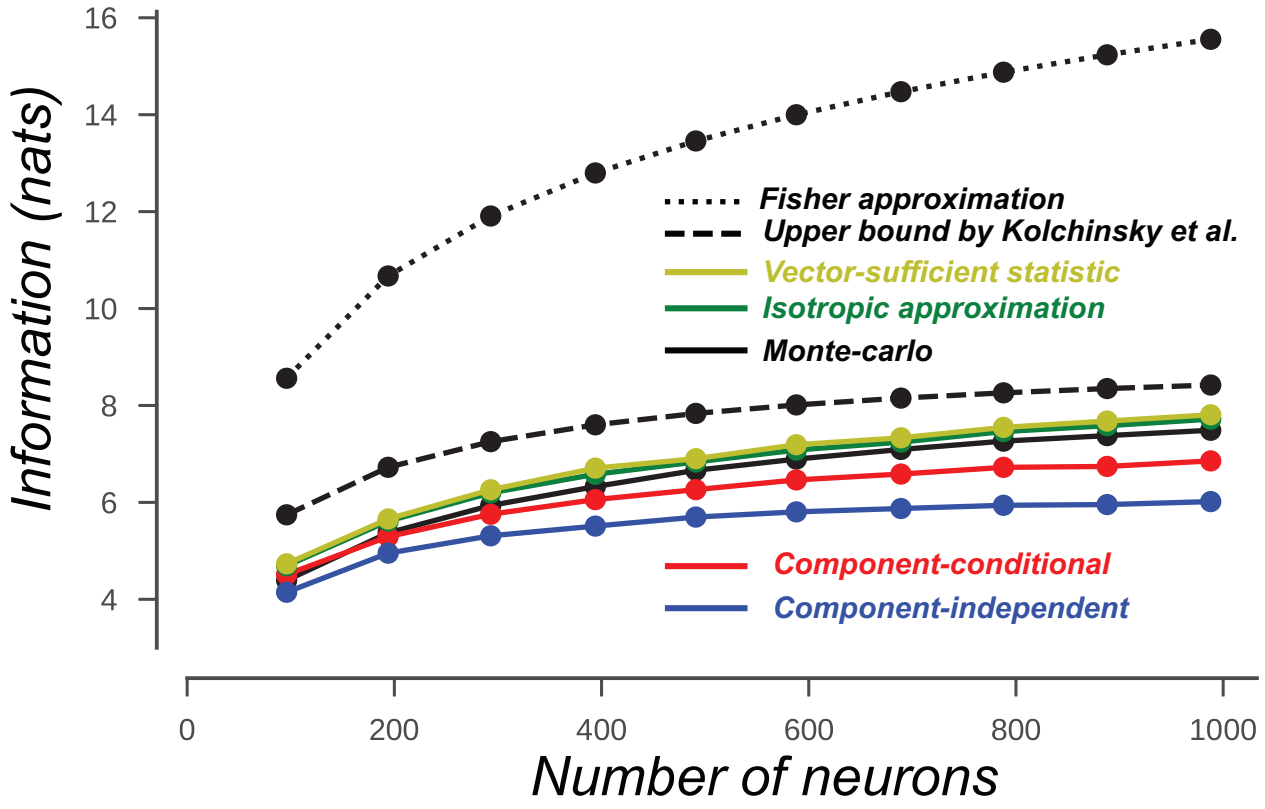

Figure 3.17: **Information curves for the population of section 3.18.1 compared to Fisher approximation**. Lines and errorbars are mean and standard deviation over ten repeats of the estimator. $\hat{I}(\vec{R},\vec{S})$ (solid black line), $I_{\text{k-w}}(\vec{R},\vec{S})$ (dashed black line), $I_{\text{vector}}(\vec{S},\vec{T})$ (solid yellow line), $I_{\text{fisher}}(\vec{R},\vec{S})$ (dotted black line), $I_{\text{comp-cond}}(\vec{S},\vec{T})$ (solid red line), $I_{\text{comp-ind}}(\vec{S},\vec{T})$ (solid blue line), $I_{\text{iso}}(\vec{S},\vec{T})$ (solid green line)

3.24 Appendix E: Extension to polynomial activation functions

In section 3.17.1 we assumed that the activation functions were affine functions of the stimulus vector $\vec{s}$. In this appendix we will show how to generalize some of the results of section 3.17.1 to polynomial activation functions. For clarity of exposition we will demonstrate this generalization for quadratic functions as the procedure for higher order polynomials follows quickly. To begin, we add a quadratic term to Eq (3.80):

$$f_n(\vec{s}) = \vec{s}^T \gamma_n \vec{s} + \vec{w}_n \cdot \vec{s} - \alpha_n \tag{3.146}$$

$\gamma_n \in \mathcal{R}^{D \times D}$ is a symmetric $D \times D$ matrix representing the quadratic kernel of the n^{th} neuron's activation function. We note that $\vec{s}^T \gamma_n \vec{s} = \vec{s}\,\vec{s}^T \circ \gamma_n$ where $A \circ B$ is the hadamard product between equally shaped matrices A and B and $\vec{s}\,\vec{s}^T \in \mathcal{R}^{D \times D}$ is the outer product of $\vec{s}$ with itself. We define a vector embedding of $\vec{s}$ into $\mathcal{R}^{D+D^2}$, $\vec{\psi}(\vec{s})$:

$$\psi(\vec{s})_d = \begin{cases} s_d & \text{if } d \leq D \\ s_a * s_b & \text{if } d > D \end{cases}$$

Where $a = d \mod D$ and $b = \lfloor d/D \rfloor$ are index mappings that map a $D \times D$ matrix into a vector of length D^2. We define a similar vector embedding of $\vec{w}_n$ and γ_n, $\vec{\tau}_n(\vec{w}_n, \gamma_n)$:

$$\tau_n(\vec{w}_n, \gamma_n)_d = \begin{cases} w_{n,d} & \text{if } d \leq D \\ \gamma_{n,ab} & \text{if } d > D \end{cases}$$

For the sake of brevity we henceforth omit the dependence of $\vec{\tau}_n$ on $\vec{w}_n$ and γ_n. By construction we have the following equivalence:

$$\vec{s}^T \gamma_n \vec{s} + \vec{w}_n \cdot \vec{s} = \vec{\tau}_n \cdot \vec{\psi}(\vec{s}) \tag{3.147}$$

If all neurons have activation functions of the form in (3.146) then $P(\vec{r}|\vec{s})$ may once again be written as an exponential family

$$\begin{aligned} P(\vec{r}|\vec{s}) &= h(\vec{r}) \exp(\vec{t}^{\,\mathrm{quad}}(\vec{r}) \cdot \vec{\psi}(\vec{s}) - A(\vec{s})) \\ A(\vec{s}) &= \sum_n \log(2\cosh(\vec{\tau}_n \cdot \vec{\psi}(\vec{s}) - \alpha_n)) \\ \vec{t}^{\,\mathrm{quad}}(\vec{r}) &= \sum_n \vec{\tau}_n r_n \end{aligned} \tag{3.148}$$

As $\vec{t}^{\,\mathrm{quad}}(\vec{r}) \in \mathcal{R}^{D+D^2}$ is the sufficient statistic for this family, and $\vec{\psi}(\vec{s})$ is the natural parameter. As before, $I(\vec{R}, \vec{S}) = I(\vec{T}^{\,\mathrm{quad}}, \vec{S})$. However, we note that $P(\vec{r}|\vec{S})$ can be written entirely in terms of $\vec{\psi}$. Additionally, we note that the support of $\vec{\psi}$ lies on a D-dimensional manifold in $\mathcal{R}^{D+D^2}$ and $\vec{s}$ maps injectively into this manifold. Thus $I(\vec{T}^{\,\mathrm{quad}}, \vec{S}) = I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi})$.

We note several properties of $I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi})$. First, we can in principle expand $I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi})$ like Eq. (3.87):

$$I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi}) = \sum_{d=1}^{d=D+D^2} I(\vec{T}^{\,\mathrm{quad}}, \Psi_d|\Psi_{<d}) \tag{3.149}$$

Secondly, the same reduction as Eq. (3.94) holds for $I(\vec{T}^{\,\mathrm{quad}}, \Psi_d|\Psi_{<d})$:

$$I(\vec{T}^{\,\mathrm{quad}}, \Psi_d|\Psi_{<d}) = I(T^{\mathrm{quad}}_{\geq d}, \Psi_d|\Psi_{<d}) \tag{3.150}$$

Most notably however, is that $I(\vec{T}^{\,\mathrm{quad}}, \Psi_d|\Psi_{<d}) = I(T^{\mathrm{quad}}_{\geq d}, \Psi_d|\Psi_{<d}) = 0$ for $d > D$. This holds

because $\psi_d = g(\psi_{\leq D})$ for $d > D$, where $g(\psi_{\leq D})$ is just the product of the two relevant components of $\psi_{\leq D}$. Because of this functional dependence we can just apply the generalization of Eq. (3.90). Therefore the expansion of $I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi}\,)$ can be truncated after D terms.

$$I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi}\,) = \sum_{d=1}^{d=D} I(T^2_{\geq d}, \Psi_d | \Psi_{<d}) = \sum_{d=1}^{d=D} I(T^{\mathrm{quad}}_{\geq d}, S_d | S_{<d}) \tag{3.151}$$

In fact, we can make an even stronger reduction by noting that conditioning on components of $\vec{S} = \psi_{\leq D}$ effectively also conditions on elements of $\psi_{>D}$. For clarity of exposition we break down $\vec{T}^{\,\mathrm{quad}}$ into vector and matrix valued components:

$$\begin{aligned}
\vec{T}^{\,\mathrm{quad}} &\equiv \left\{ T_{\leq D}, T^{\leq D}_{\leq D} \right\} \\
T^{\mathrm{quad}}_{\leq D} &\rightarrow T_{\leq D} \in \mathcal{R}^D \\
T^{\mathrm{quad}}_{>D} &\rightarrow T^{\leq D}_{\leq D} \in \mathcal{R}^{D \times D}
\end{aligned}$$

We note that conditioning on S_d conditions on the components of $\vec{\psi}$ corresponding to T_d and T^d_d. Additionally conditioning on $S_{<d}$ conditions on the components of $T_{<d}$ *and* on the components of $T^{d_2}_{d_1}$ for all indices d_1 and d_2 such that $1 \leq d_1, d_2 < d$. Thus Eq. (3.151) can be further generalized:

$$I(\vec{T}^{\,\mathrm{quad}}, \vec{S}\,) = \sum_{d=1}^{d=D} I\left(\left\{ T_{\geq d}, T^{\geq d}_{\geq d} \right\}, S_d \middle| S_{<d} \right) \tag{3.152}$$

The d^{th} term in Eq (3.151) has $D^2 + D + 1$ degrees of freedom while the d^{th} term in Eq (3.152) has $D^2 + D + 1 - (d-1)^2$ degrees of freedom. The above procedure can be generalized to polynomial activation functions of arbitrarily high but finite order, though the dimensionality of the sufficient statistic and natural parameter grow exponentially with the order. However, Eq. (3.152) holds for any order of polynomial, so that one one needs only compute the first D terms of the expansion of

the mutual information between the sufficient statistic and natural parameter.

3.25 Acknowledgments

Chapter 3, in full, is a reformatted version of the material as it appears in the following publication: John A. Berkowitz, Tatyana O. Sharpee. Quantifying information conveyed by large neuronal populations. Neural Computation, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 4: Conclusion

In this thesis we have taken a broad look at issues in identifying optimal representations of neural population responses from an information theoretic point of view. While there may be many representations that preserve all information in the full population response, a good representation should also be biologically plausible and easily decodable. This constraint has also been highlighted in recent work on representation learning in deep neural networks [Amjad and Geiger, 2018]. The success of finding an optimal representation is dependent upon the ability the evaluate a representation's performance, e.g. the mutual information. Thus, in addition to studying the properties of our proposed representation, we also examined the task of estimating the amount of information transmitted by such a population.

In chapter 2 we showed that a broad class of models of spiking neural populations yield a sufficient statistic with dimensionality independent of the population size. This result hold true even in the case of certain types of intrinsic correlation between neurons. The key requirement underlying this result is that the neural population follows a maximum entropy distribution, a common assumption in many analyses of neural data [Granot-Atedgi et al., 2013]. Additionally we showed that, under certain conditions on the receptive field distribution, the sufficient statistic itself can serve as a decoder of the stimulus subspace, meriting its use as a representation of the population code compared to other information preserving transformations.

There are several natural directions of further inquiry for this work. The decoder we proposed in chapter 2 produces a point estimate, that is it returns a single value for any given value of the sufficient statistic. An increasingly popular framework for understanding neural codes is probabilistic inference, where the neural response is used to infer properties of the posterior distribution over the stimulus, either implicitly or explicitly. Knowledge of the posterior distribution can be used to derive various point estimates and quantify uncertainty in the stimulus estimate, and there is evidence that uncertainty quantification occurs during decision making (citation here). One popular approach to probabilistic inference includes treating the posterior density or, equivalently, the likelihood profile as a linear combination of basis functions with weights determined by firing rates of individual neurons [Jazayeri and Movshon, 2006]. Exponential families are well suited for this approach. An alternative approach, the sampling hypothesis, focuses on transforming samples from the population response into samples from the posterior. We believe that variations on the decoder presented in chapter 2 could be used to approximate such a response-posterior transformation.

In chapter 3 we explored approaches to estimating the information transmitted by models in the same class as those considered in the previous chapter. We showed that decompositions based upon the chain rule for mutual information combine synergistically with the properties of exponential families to yield efficient estimators and lower bounds of the generally intractable mutual information. In particular, decompositions over the corresponding coordinates of the stimulus and sufficient statistic provide a tractable lower bound that compares favorably to other non-parametric approaches to estimating the mutual information.

While we chose to approximate the decomposition in terms of pairs of scalar variables for computational expediency, there are other motivations for studying decompositions of the mutual information between two multicomponent variables. The partial information decomposition is an approach to characterizing the semantic content of transmitted information by decomposing the

mutual information between the full stimulus and response into a series of terms corresponding to the mutual information between different subsets of the stimulus and response [Kay et al., 2017]. The PID also suffers from the same tractability issues as the computation of the full information, but we conjecture that the aforementioned properties of exponential families can be used to derive accurate and tractable approximations of the PID. Additionally, though we found the ordering heuristic presented in chapter 3 to perform well for low dimensions, there may be more principally motivated orderings based upon approximating $P(\vec{t}|\vec{s})$ by a graphical model of bounded degree [Chow and Liu, 1968].

Bibliography

[Agakov and Barber, 2004] Agakov, F. and Barber, D. (2004). Variational information maximization for neural coding. In *Neural Information Processing*, pages 543–548. Springer.

[Amjad and Geiger, 2018] Amjad, R. A. and Geiger, B. C. (2018). How (not) to train your neural network using the information bottleneck principle. *arXiv preprint arXiv:1802.09766*.

[Atick and Redlich, 1990] Atick, J. J. and Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Comput.*, 2:308–320.

[Banerjee et al., 2005] Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749.

[Barber and Agakov, 2003] Barber, D. and Agakov, F. (2003). The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 201–208. MIT Press.

[Beck et al., 2008] Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, 60:1142–1152.

[Belghazi et al., 2018] Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. (2018). Mine: Mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

[Bell and Sejnowski, 1997] Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.*, 23:3327–3338.

[Bender and Orszag, 1999] Bender, C. and Orszag, S. (1999). *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Advanced Mathematical Methods for Scientists and Engineers. Springer.

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

[Berkowitz and Sharpee, 2018] Berkowitz, J. and Sharpee, T. (2018). Decoding neural responses with minimal information loss. *bioRxiv*.

[Bialek, 2012] Bialek, W. (2012). *Biophysics: Searching for principles*. Princeton University Press, Princeton and Oxford.

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

[Brenner et al., 2000b] Brenner, N., Strong, S. P., Koberle, R., Bialek, W., and de Ruyter van Steveninck, R. R. (2000b). Synergy in a neural code. *Neural Comput*, 12:1531–1552.

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

[Candès and Wakin, 2008] Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30.

[Carandini and Heeger, 2011] Carandini, M. and Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62.

[Chow and Liu, 1968] Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.

[Cohen and Maunsell, 2009] Cohen, M. R. and Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–1600.

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

[Dasgupta et al., 2017] Dasgupta, S., Stevens, C. F., and Navlakha, S. (2017). A neural algorithm for a fundamental computing problem. *Science*, 358(6364):793–796.

[Deserno, 2004] Deserno, M. (2004). How to generate equidistributed points on the surface of a sphere. *P.-If Polymerforshung (Ed.)*, page 99.

[Dettner et al., 2016] Dettner, A., Münzberg, S., and Tchumatchenko, T. (2016). Temporal pairwise spike correlations fully capture single-neuron information. *Nature communications*, 7:13805.

[Ecker et al., 2011] Ecker, A. S., Berens, P., Tolias, A. S., and Bethge, M. (2011). The effect of noise correlations in populations of diversely tuned neurons. *Journal of Neuroscience*, 31(40):14272–14283.

[Fairhall et al., 2001] Fairhall, A. L., Lewen, G. D., Bialek, W., and de Ruyter van Steveninck, R. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, pages 787–792.

[Gao et al., 2015] Gao, S., Ver Steeg, G., and Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables. In *Artificial Intelligence and Statistics*, pages 277–286.

[Gao et al., 2017] Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017). Estimating mutual information for discrete-continuous mixtures. In *Advances in Neural Information Processing Systems*, pages 5986–5997.

[Gao et al., 2018] Gao, W., Oh, S., and Viswanath, P. (2018). Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*.

[Georgopoulos et al., 1986] Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233:1416–1419.

[Granot-Atedgi et al., 2013] Granot-Atedgi, E., Tkacik, G., Segev, R., and Schneidman, E. (2013). Stimulus-dependent maximum entropy models of neural population activity. *Plos Comp Biol*, 9:e1002922.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning*, pages 9–41. Springer.

[Haussler et al., 1997] Haussler, D., Opper, M., et al. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492.

[Hohl et al., 2013] Hohl, S. S., Chaisanguanthum, K. S., and Lisberger, S. G. (2013). Sensory population decoding for visually guided movements. *Neuron*, 79:167–179.

[Hosoya et al., 2005] Hosoya, T., Baccus, S. A., and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, 436:71–77.

[Huang and Zhang, 2018] Huang, W. and Zhang, K. (2018). Information-theoretic bounds and approximations in neural population coding. *Neural computation*, (Early Access):1–60.

[Huang and Lisberger, 2009] Huang, X. and Lisberger, S. G. (2009). Noise correlations in cortical area mt and their potential impact on trial-by-trial variation in the direction and speed of smooth-pursuit eye movements. *Journal of Neurophysiology*, 101(6):3012–3030.

[Hyvärinen et al., 2009] Hyvärinen, A., Hurri, J., and Hoyer, P. O. (2009). Natural image statistics.

[Janzamin et al., 2014] Janzamin, M., Sedghi, H., and Anandkumar, A. (2014). Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*.

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

[Josić et al., 2009] Josić, K., Shea-Brown, E., Doiron, B., and de la Rocha, J. (2009). Stimulus-dependent correlations and population codes. *Neural computation*, 21(10):2774–2804.

[Kardar, 2007] Kardar, M. (2007). *Statistical physics of fields*. Cambridge University Press.

[Kastner and Baccus, 2011] Kastner, D. B. and Baccus, S. A. (2011). Coordinated dynamics encoding in the retina using opposing forms of plasticity. *Nature Neurosci.*, 14(10):1317–1322.

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

[Kay et al., 2017] Kay, J. W., Ince, R. A., Dering, B., and Phillips, W. A. (2017). Partial and entropic information decompositions of a neuronal modulatory interaction. *Entropy*, 19(11):560.

[Kohn and Smith, 2005] Kohn, A. and Smith, M. A. (2005). Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J Neurosci.*, 25:3661–3673.

[Kolchinsky et al., 2017] Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. (2017). Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436*.

[Kraskov et al., 2004] Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6):066138.

[Laughlin et al., 1998] Laughlin, S. B., de Ruyet van Steveninck, R. R., and Anderson, J. C. (1998). The metabolic cost of neural computation. *Nat. Neurosci.*, 41:36–41.

[Ma et al., 2006] Ma, W. J., Beck, J., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neurosci.*, 9:1432–1438.

[McDonnel and Stocks, 2008] McDonnel, M. D. and Stocks, N. (2008). Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Phys. Rev. Lett.*, 101:058103.

[McDonnell et al., 2006] McDonnell, M. D., Stocks, N. G., Pearce, C. E., and Abbott, D. (2006). Optimal information transmission in nonlinear arrays through suprathreshold stochastic resonance. *Physics Letters A*, 352(3):183–189.

[Moreno-Bote et al., 2014] Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417.

[Nemenman et al., 2004] Nemenman, I., Bialek, W., and van Steveninck, R. d. R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111.

[Nemenman et al., 2002] Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. In *Advances in neural information processing systems*, pages 471–478.

[Olshausen and Field, 2004] Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr Opin Neurobiol*, 14(4):481–487.

[Opper et al., 2001] Opper, M., Winther, O., et al. (2001). From naive mean field theory to the tap equations. *Advanced mean field methods: theory and practice*, pages 7–20.

[Orban et al., 2016] Orban, G., Berkes, P., Fiser, J., and Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 483:47–52.

[Osborne et al., 2008a] Osborne, L. C., Palmer, S. E., G., L. S., and Bialek, W. (2008a). The neural basis for combinatorial coding in a cortical population response. *J Neurosci*, 28:13522–13531.

[Osborne et al., 2008b] Osborne, L. C., Palmer, S. E., G., L. S., and Bialek, W. (2008b). The neural basis for combinatorial coding in a cortical population response. *J Neurosci*, 28:13522–13531.

[Paninski, 2003] Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

[Ratliff et al., 2010] Ratliff, C. P., Borghuis, B. G., Kao, Y.-H., Sterling, P., and Balasubramanian, V. (2010). Retina is structured to process an excess of darkness in natural scenes. *Proceedings of the National Academy of Sciences*, 107(40):17368–17373.

[Reich et al., 2001] Reich, D. S., Mechler, F., and Victor, J. (2001). Independent and redundant information in nearby cortical neurons. *Science*, 294:2566–2568.

[Renner and Maurer, 2002] Renner, R. and Maurer, U. (2002). About the mutual (conditional) information. In *Proc. IEEE ISIT*.

[Rieke et al., 1997] Rieke, F., Warland, D., de Ruyter van Steveninck, R. R., and Bialek, W. (1997). *Spikes: Exploring the neural code*. MIT Press, Cambridge.

[Rockafellar and Wets, 2009] Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

[Salinas and Abbott, 1994] Salinas, E. and Abbott, L. F. (1994). Vector reconstruction from firing rates. *J Comp Neurosci*, 1:89–107.

[Schneidman et al., 2006] Schneidman, E., Berry II, M. J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007=1012.

[Schneidman et al., 2003] Schneidman, E., Bialek, W., and Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553.

[Schneidman et al., 2001] Schneidman, E., Slonim, N., Tishby, N., de Ruyter van Stevenick, R. R., and Bialek, W. (2001). Analyzing neural code using the information bottleneck method. In *Advances in Neural Information Processing*, volume 14.

[Sharpee et al., 2008a] Sharpee, T. O., Miller, K. D., and Stryker, M. P. (2008a). On the importance of static nonlinearity in estimating spatiotemporal neural filters with natural stimuli. *J. Neurophysiol.*, 99:2496–2509.

[Sharpee et al., 2008b] Sharpee, T. O., Stryker, M. P., and Miller, K. D. (2008b). On the importance of the static nonlinearity in estimating spatiotemporal neural filters with natural stimuli. *J. of Neurophys.*, 99:2496–2509.

[Sharpee et al., 2006a] Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., and Miller, K. D. (2006a). Adaptive filtering enhances information transmission in visual cortex. *Nature*, 439:936–942.

[Sharpee et al., 2006b] Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., and Miller, K. D. (2006b). Adaptive filtering enhances information transmission in visual cortex. *Nature*, 439(7079):936–942.

[Smith and Lewicki, 2006] Smith, E. and Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, 439:978–82.

[Sompolinsky et al., 2001] Sompolinsky, H., Yoon, H., Kang, K., and Shamir, M. (2001). Population coding in neuronal systems with correlated noise. *Phys. Rev. E*, 61:051904.

[Srivastava et al., 2017] Srivastava, K. H., Holmes, C. M., Vellema, M., Pack, A. R., Elemans, C. P., Nemenman, I., and Sober, S. J. (2017). Motor control by precisely timed spike patterns. *Proceedings of the National Academy of Sciences*, 114:201611734.

[Stein, 1981] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151.

[Strong et al., 1998] Strong, S. P., Koberle, R., van Steveninck, R. R. d. R., and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical review letters*, 80(1):197.

[Tao and Vu, 2010] Tao, T. and Vu, V. (2010). A sharp inverse littlewood-offord theorem. *Random Structures & Algorithms*, 37(4):525–539.

[Theunissen and Miller, 1995] Theunissen, F. and Miller, J. P. (1995). Temporal encoding in nervous systems: a rigorous definition. *Journal of computational neuroscience*, 2(2):149–162.

[Tkačik et al., 2010] Tkačik, G., Prentice, J. S., Balasubramanian, V., and Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences*, 107(32):14419–14424.

[Treves and Panzeri, 1995] Treves, A. and Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Comp.*, 7:399–407.

[Wainwright, 1999] Wainwright, M. J. (1999). Visual adaptation as optimal information transmission. *Vision Res*, 39:3960–3974.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

[Yu et al., 2010] Yu, Y., Crumiller, M., Knight, B., and Kaplan, E. (2010). Estimating the amount of information carried by a neuronal population. *Frontiers in computational neuroscience*, 4:10.

[Zhang and Sharpee, 2016] Zhang, Y. and Sharpee, T. O. (2016). A robust feedforward model of the olfactory system. *PLoS Comput. Biol.*, 12:e1004850.

[Zohary et al., 1994] Zohary, E., Shadlen, M. N., and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 86:140–143.