Characterizing Redundancy in Populations of Neurons

Jiening Zhan

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-60
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-60.html

May 20, 2008

Characterizing Redundancy in Populations of Neurons

by

Jiening Zhan

B.S. Washington University in St. Louis 2005
M.E.E. University of California, Berkeley 2007

A thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Engineering - Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael Gastpar, Chair
Professor Jose Carmena

Fall 2007

The thesis of Jiening Zhan is approved.

Chair	Date
	Date

University of California, Berkeley

Fall 2007

Characterizing Redundancy in Populations of Neurons

by

Jiening Zhan

Abstract

Characterizing Redundancy in Populations of Neurons

by

Jiening Zhan

Master of Science in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Michael Gastpar, Chair

Populations of neurons often respond in a redundant fashion to stimuli. One such redundancy is that multiple neurons react to similar stimulus features, and another is that neurons excite or inhibit each other. Using information theoretic principles, redundancy measures are defined and analyzed for a series of theoretical models in this thesis. Furthermore, the redundancy measures are applied to measurement data from populations of neurons in the auditory system of Zebra Finch Song Birds. The data suggests that the amount of redundancy varies from one population to another. This thesis attempts to understand and to distinguish between neural populations based on their relative amounts of redundancies. Finally, a coarse approach to characterizing redundancy is developed via considering the mutual information between the stimulus and the responses of one, two, three etc. neurons in the considered population. This coarse characterization is analyzed for several theoretical models. The results show that a rapid increase in information with the number of neurons in the population suggests high redundancy between neurons while a slow increase implies low redundancy.

Professor Michael Gastpar
Thesis Committee Chair

Contents

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

I would like to thank my adviser, Professor Michael Gastpar. With his enthusiasm, his inspiration, and his efforts to explain things clearly, he helped make this research fun for me. Throughout my thesis-writing period, he provided encouragement, sound advice, good teaching, and lots of good ideas.

I would like to thank everyone who collaborated with this project, including Professor Frederic Theunissen, Mauro Merolle, and Patrick Gill. They provided valuable insight about this project from a neuroscience perspective.

I would like to thank Professor Jose Carmena for being the reader of this thesis and for giving many helpful comments.

Finally, I would like to thank Anand Sarwate for his help in technical aspects of this thesis.

The work in this thesis was supported in part by NIDCD/NIH under Grant Number 1R01DC007293-01 'CRCNS: Ethological theories: optimal auditory processing' and by an NSF Graduate Fellowship.

Curriculum Vitæ

Jiening Zhan

Education

2005	Washington University in St. Louis
	B.S., Electrical Engineering

Personal

Born	May 3, 1985, Guangzhou, China

Chapter 1

Neuroscience and Redundancy: An Introduction

1.1 Redundancy in Population Coding

When presented with a sensory stimuli, a neuron's response is in the form of an action potential, often called a 'spike'. The representation of information in the neuron's spiking response, known as neural coding, has been (and continues to be) an area of intense investigation in theoretical neuroscience [3], [10]. However, although a single neuron has perceptible influence on the animal's behavior, it is ultimately large groups of neurons that have significant impact on behavior [6], [7]. One interesting question that arises is the mystery behind population coding - the manner in which groups of neurons interact to represent information about the sensory stimuli. Schneidman et. al [9] suggests that ensembles of neurons can encode information in either a synergistic, redundant, or independent fashion. That is, an ensemble of neurons can encode more, less, or same information than the sum of each individual neuron in the ensemble.

The idea of population coding can be elucidated by abstracting the neural system and considering a mathematical model representation as shown in figure 1.1. The model suggest that at a given time t, $S(t)$ represents the stimulus presented to the M neurons in the

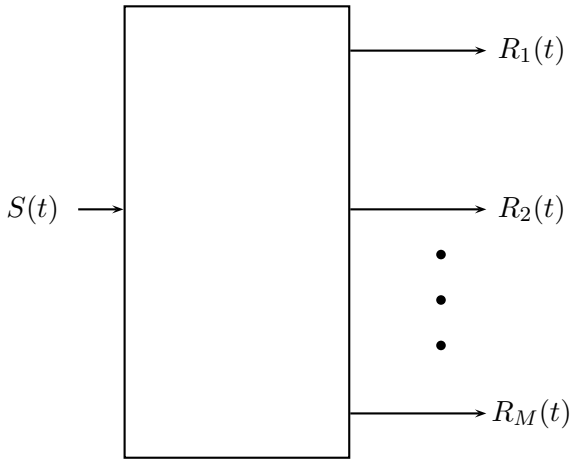


Figure 1.1. Abstraction of the Neural System.

population and $R_i(t)$ represents the response of ith neuron. Basic insight and past literature suggests that there exists an inherent degree of redundancy in the M neural reponses $R_1(t), R_2(t), ..., R_M(t)$. The fundamental question that arises is how to define and characterize this redundancy. More generally, given a population of neurons, how is the redundancy of the population measured? The answer to this question will enable the characterization of a neural region based on its redundancy and the distinction of different neural regions based on their respective redundancies. This will give insight into the general representation and transmission of information in the brain.

1.2 Notation and Basic Definitions

The first part of this section introduces the notation used in this thesis. The second part provides definitions and properties of elementary probabilitistic and information theoretic measures.

1.2.1 Notation

- S: stimulus

- M: number of neurons in the population
- R_i: response of neuron i (for $i \in 1, ..., M$)
- W_i: noise in response i
- $\mathbf{R}$: population response $(R_1, ..., R_M)$

1.2.2 Basic Definitions

Definition 1. *The sample space Ω of an experiment or random trial is the set of all possible outcomes*

Example 1. *For a coin toss, the two possible outcomes of the experiment are heads and tails. Letting H denote heads and T denote tails, it follows that the sample space $\Omega = \{H, T\}$*

Definition 2. *A random variable X is a real valued function defined over a sample space Ω.*

In mathematics, $X : \Omega \to \mathbb{R}$. There are two types of random variables: discrete and continuous. The discrete random variables takes on countable number of possible values, while a continuous random variable takes on uncountable number of possible values.

Definition 3. *For a discrete random variable X, the associated probability distribution $p(x)$ is a function that gives the probability that X takes on the value x.*

Definition 4. *For a continuous random variable X, the probability density function $f(x)$ is defined to be such that given any interval $[a, b]$, the probability that X takes a value on $[a, b]$ is given by*

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{1.1}$$

Definition 5. *The expectation $E[X]$ of a random variable X is defined as*

$$E[X] = \sum_x xp(x) \tag{1.2}$$

for X discrete, and

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \tag{1.3}$$

for X continuous.

Definition 6. *The Covariance $Cov(X)$ of a random variable X is defined as*

$$Cov(X) = E\left[X^2\right] - E^2[X] \tag{1.4}$$

Note that $Cov(X) \geq 0$.

Definition 7. *A bernoulli random variable X with parameter p takes values in $\{0, 1\}$ with $P(X = 1) = p$ and $P(X = 0) = 1 - p$. $X \sim \mathcal{B}(p)$ can be written to describe X.*

Definition 8. *A gaussian random variable X can be defined by two paramters: mean and covariance. If X is gaussian with mean μ and covariance σ^2, then X has density*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \tag{1.5}$$

for each $x \in \mathbb{R}$. $X \sim \mathcal{N}(\mu, \sigma^2)$ can be written to describe X.

Definition 9. *The entropy $H(X)$ of a discrete random variable X with probability distribution $p(x)$ is defined as*

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{1.6}$$

while the entropy $h(X)$ for a continuous random variable X with probability density $f(x)$ is defined as

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx \tag{1.7}$$

Definition 10. *For $X \sim \mathcal{N}(\mu, \sigma^2)$, the entropy $h(X)$ is given by*

$$h(X) = \frac{1}{2} \log 2\pi e\sigma^2 \tag{1.8}$$

Definition 11. *The joint entropy $H(X_1, X_2, ..., X_M)$ of a set of discrete random variables $X_1, X_2, ..., X_M$ with joint distribution $p(x_1, x_2, ..., x_M)$ is defined as*

$$H(X_1, X_2, ..., X_M) = -\sum_{x_1,...,x_M} p(x_1, x_2, ..., x_M) \log p(x_1, x_2, ..., x_M) \tag{1.9}$$

The joint entropy $h(X_1, X_2, ..., X_M)$ of a set of continuous random variables $X_1, X_2, ..., X_M$ with joint density $f(x_1, x_2, ..., x_M)$ is defined as

$$h(X_1, X_2, ..., X_M) = -\int f(x_1, x_2, ..., x_M) \log f(x_1, x_2, ..., x_M) dx_1, dx_2, ...dx_M \tag{1.10}$$

Definition 12. *The conditional entropy $H(X_1|X_2)$ of discrete random variables X_1, X_2 with joint distribution $p(x_1, x_2)$ is defined as*

$$H(X_1|X_2) = -\sum_{x_1,x_2} p(x_1, x_2) \log p(x_1|x_2) \tag{1.11}$$

The joint entropy $h(X_1, X_2)$ of a set of continuous random variables X_1, X_2 with joint density $f(x_1, x_2)$ is defined as

$$h(X_1|X_2) = -\int f(x_1, x_2, ..., x_M) \log f(x_1|x_2) dx_1, dx_2 \tag{1.12}$$

Theorem 1. *For any random variables X, Y,*

$$H(X, Y) = H(X) + H(Y|X) \tag{1.13}$$

Theorem 2. *Given random variables $X_1, X_2, ..., X_M$, the chain rule of entropy states that*

$$H(X_1, X_2, ..., X_M) = \sum_{i=1}^{M} H(X_i|X^{i-1}) \tag{1.14}$$

Definition 13. *The mutual information $I(X; Y)$ between random variables X, Y is given by*

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{1.15}$$

for X, Y discrete, and

$$I(X; Y) = \sum_{x,y} f(x, y) \log \frac{f(x, y)}{f(x)f(y)} \tag{1.16}$$

for X, Y continuous.

From the above definitions, it can be seen that $I(X; Y) = I(Y; X)$ and $I(X; Y) = H(Y) - H(Y|X)$. Similar properties hold for the continuous version.

Theorem 3. *Given random variables $X, X_1, X_2, ..., X_M$, the chain rule of mutual information states that*

$$I(X; X_1; X_2; ...; X_M) = \sum_{i=1}^{M} I(X; X_i|X^{i-1}) \tag{1.17}$$

Definition 14. *The joint mutual information of a set of random variables* $X_1, ..., X_M$ *is defined as*

$$I(X_1; X_2, ...; X_M) = \sum_{i=1}^{M} H(X_i) - H(X_1, X_2, ..., X_M) \tag{1.18}$$

Similar definition holds for continuous version.

Definition 15. *The KL divergence* $D(p(x)||q(x))$ *between probability distributions* $p(x)$ *and* $q(x)$ *is defined as*

$$D(p(x)||q(x)) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \tag{1.19}$$

1.3 Measures of Redundancy

In the sixties, Glovazky [1] explored the concept of redundancy in a set of patterns, which is any group of geometrical configurations which obey a set of conditions. In his work, any of the given patterns can be divided into a finite number of cells. Therefore, each pattern can be identified based on the contents of all its cells. However, depending on the given pattern set, a pattern may be identified without exploring the contents of all its cells since the information of certain cells may be redundant. In his paper, Glovazky devised a method by which the cells that are redundant with respect to the given identification process can be determined.

Glovazky's concept of redundancy in patterns influences the idea of redundancy in population coding. Consider a neural ensemble with M neurons with population response $R_1, ..., R_M$ to sensory stimulus S. Assume the neurons encode information in a redundant fashion. For a given $N \leq M$, let $R_1, ..., R_N$ be a subset of $R_1, ..., R_M$. The encoding in the neural system could be performed such that most of the information about the stimulus can be determined from $R_1, ..., R_N$. Therefore, R_i for all $i > N$ does not give much additional information not already present in $R_1, ..., R_N$. Thus, the responses R_i for all $i > N$ are redundant.

In regards to population coding specifically, a few redundancy measures have been found in literature. Chechik et. al [4] defined a redundancy measure based on the joint mutual

information of the population response. Given M neurons that output responses $R_1, ..., R_M$ when being exposed to stimulus S, the population redundancy r is defined as

Definition 16.

$$r = \frac{I(R_1; R_2; ...; R_M)}{\sum_{i=1}^{M} I(S; R_i)} \tag{1.20}$$

A neural population in which the information in each response is completely different from the information in any other response is termed independent. Conversely, a neural population in which the information in each response is the same as every other response is called fully dependent. Intuitively, an independent population should have redundancy zero while a fully dependent population should have redundancy one. Therefore, a well defined redundancy measure should lie within the $[0, 1]$ interval. Following are properties for the redundancy measure r from definition 16 and conditions for it to lie within the desire $[0, 1]$ interval.

Property 1. *$r \geq 0$ with equality iff $R_1, ..., R_M$ independent*

Proof. Using the chain rule of entropy and the fact that conditioning reduces entropy, it follows that

$$H(R_1, ..., R_M) = \sum_{i=1}^{M} H(R_i | R_{i-1}, ..., R_1) \tag{1.21}$$

$$\leq \sum_{i=1}^{M} H(R_i) \tag{1.22}$$

Equality is achieved in the above iff $R_1, ..., R_M$ are independent. This implies that $r \geq 0$ with equality iff $R_1, ..., R_M$ are independent. □

From definition 16, it can be inferred that correlation between the responses $R_1, ..., R_M$ increases the joint information $I(R_1; R_2; ...; R_M)$ and thus causes redundancy in the population. This aligns with the basic intuition that redundancy is the repetition of information.

Property 2. *Given an unordered population response $\mathbf{R}$, if there exists an ordering of $\mathbf{R}$ such that $R_1, R_2, ..., R_M$ satisfies $\sum_{i=1}^{M} I(R_i; S) \geq \sum_{i=2}^{M} I(R_i; R^{i-1})$, then $r \leq 1$.*

Proof. The result is simply seen by the definition of mutual information

$$\sum_{i=2}^{M} I(R_i; R^{i-1}) = \sum_{i=2}^{M} H(R_i) - H(R_i|R^{i-1}) \tag{1.23}$$

$$= \sum_{i=1}^{M} H(R_i) - H(R_1, R_2, ..., R_M) \tag{1.24}$$

□

Property 2 gives a condition for the population redundancy r to be within the $[0, 1]$, the desired interval.

Another definition of redundancy in population coding comes from Reich et. al [8]. Their measure, given below, is based on the interaction between pairwise neurons.

$$\Delta_{12} = \frac{I(R_1; S) + I(R_2; S) - I(R_1, R_2; S)}{\min I(R_1; S), I(R_2; S)} \tag{1.25}$$

This measure of redundancy depends on the difference between the information the pairwise neurons contain about the stimulus $I(S; R_1, R_2)$ and the sum of the information each individual neuron contains about the stimulus $I(S; R_1) + I(S; R_2)$. An extension of this redundancy measure to multiple neurons can be made. Let the extended redundancy measure be denoted by r'. It is given as

Definition 17.

$$r' = \frac{\sum_{i=1}^{M} I(S; R_i) - I(S; R_1, R_2, ..., R_M)}{\sum_{i=1}^{M} I(S; R_i)} \tag{1.26}$$

From this measure, it is seen that characterization of population redundancy can be accomplished by solely considering the quantity $I(S; R_1, ..., R_M)$. Note that $\sum_{i=1}^{M} I(S; R_i)$ acts merely as a normalizing term that ensure $r' \leq 1$. Calculating $I(S; R_1, R_2, ..., R_M)$ is an elusive task that involves estimating the probability distribution $p(R_1, ...R_M|S)$. However, neural data is limited, and a decent estimation of $p(R_1, ...R_M|S)$ requires an excess amount of data. Therefore, methods are needed to overcome this elusive task. This thesis proposes two such methods: a direct method and a redundancy scaling method. In the first method, a direct measure of redundancy is obtained using an estimation of the joint

entropy $H(R_1, R_2, ...R_M)$. In the second approach, based on a model of the neural system, a scaling redundancy measure is obtained.

Property 3. *If $R_1, R_2, ..., R_M$ are conditionally independent given S, then $r = r'$.*

Proof. The denominators of r and r' are the same. Using the definition of mutual information, the numerator of r' can be rewritten as

$$\sum_{i=1}^{M}(H(R_i) - H(R_i|S)) - H(R_1, ..., R_M) + H(R_1, ..., R_M|S) \tag{1.27}$$

Using the fact that $R_1, R_2, ..., R_M$ are conditionally independent given S, it follows that

$$H(R_1, ..., R_M|S) = \sum_{i=1}^{M} H(R_i|S) \tag{1.28}$$

Using the above, the numerator of r' can be simplified to

$$\sum_{i=1}^{M} H(R_i) - H(R_1, ..., R_M) \tag{1.29}$$

this is precisely the numerator of r. □

Chapter 2

Redundancy of Simple Models

The population redundancy r from definition 16 is shown to be always greater than zero. However, conditions given in property 2 must be satisfied in order for r to be less than one. In this chapter, the behavior of r is analyzed for a series of simple models. Since the entropy of gaussian random variables can be calculated easily, most of the examples considered model the stimulus and response as gaussians. These examples shows that for certain models, the redundancy r lies within $[0, 1]$, the desired interval. However, for other models, r does not have such a nice behavior.

2.1 Conditionally Independent Gaussian Model

The first example models the stimulus as a gaussian random variable, and the responses as conditionally independent gaussian random variables given the stimulus. Hence, the noise in each response is independent. In actual experiments, single unit recording is used to obtain the action potentials of neurons. Since the neural responses are measured separately, it is reasonable to assume that the noise in each response is independent.

Definition. *The stimulus* $S \sim \mathcal{N}(0, \sigma^2)$. *The noise* W_i *are independent and identically distributed (i.i.d)* $\mathcal{N}(0, \sigma_w^2)$. *Also, the* W_i *are independent of* S. *Each response* $R_i = S + W_i$.

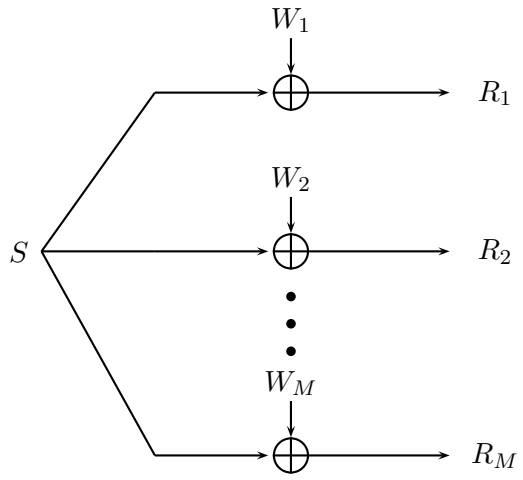


Figure 2.1. Conditionally Independent Gaussian Model.

Therefore, the population response $R_1, R_2, ..., R_M$ are gaussian random variables conditionally independent given S. Figure 2.1 illustrates this example.

Let σ_r^2 denote $Cov(R_i)$. For all i, $Cov(R_i) = Cov(S + W_i) = \sigma^2 + \sigma_w^2$. It follows that the entropy $H(R_i)$ for each response is given as

$$\begin{aligned} H(R_i) &= \frac{1}{2}\log 2\pi e\sigma_r^2 \\ &= \frac{1}{2}\log 2\pi e(\sigma_s^2 + \sigma_w^2) \end{aligned} \tag{2.1}$$

Let $\mathbf{K_r}$ denote the covariance matrix of of the random vector $\mathbf{R}$. Since $\mathbf{K_r}$ is a circulant matrix, $|\mathbf{K_r}|$ can be easily evaluated and found to be $(M\sigma_s^2 + \sigma_w^2)(\sigma_{w^2})^{M-1}$. From this, the joint entropy $H(R_1, ..., R_M)$ can be found to be

$$\begin{aligned} H(R_1, ..., R_M) &= \frac{1}{2}\log(2\pi e)^M |\mathbf{K_r}| \\ &= \frac{M}{2}\log 2\pi e + \frac{1}{2}(M\sigma_s^2 + \sigma_w^2) + \frac{M-1}{2}\log\sigma_w^2 \end{aligned} \tag{2.2}$$

Since S and R_i are gaussian random variables, the mutual information between the stimulus and each response can be found to be

$$I(S; R_i) = \frac{1}{2}\log(1 + \frac{\sigma_s^2}{\sigma_r^2}) \tag{2.3}$$

Using the measures above, the population redundancy r (from definition 16 of previous section) is given as

$$r = 1 - \frac{\log(1 + \frac{M\sigma_s^2}{\sigma_w^2})}{M \log(1 + \frac{\sigma_s^2}{\sigma_w^2})} \tag{2.4}$$

Property 4. *For the conditionally independent gaussian model, for* $M \in \mathbb{N}$,

(i) $0 \leq r \leq 1$

(ii) $\lim_{M\to\infty} r = 1$

(iii) r *is concave in* M*.*

Proof. (i) For any model, from property 1 of introduction, it is known that $r \geq 0$. The lower bound is obvious from equation 2.4.

(ii) From equation 2.4, it is shown that for large M,

$$r \sim 1 - \frac{\log M}{M} \tag{2.5}$$

Therefore, as $M \to \infty$, $r \to 1$.

(iii) see 2.*A*

□

Figure 2.2 plots r as a function of M. It confirms that r is concave in M and that $r \to 1$ as $M \to 1$.

2.2 General Conditionally Independent Model

This section generalizes conditionally independent gaussian model by removing the constraint that the stimulus and response are gaussian random variables. For this generalized model, although the redundancy r cannot be calculated precisely, interesting properties about r are presented.

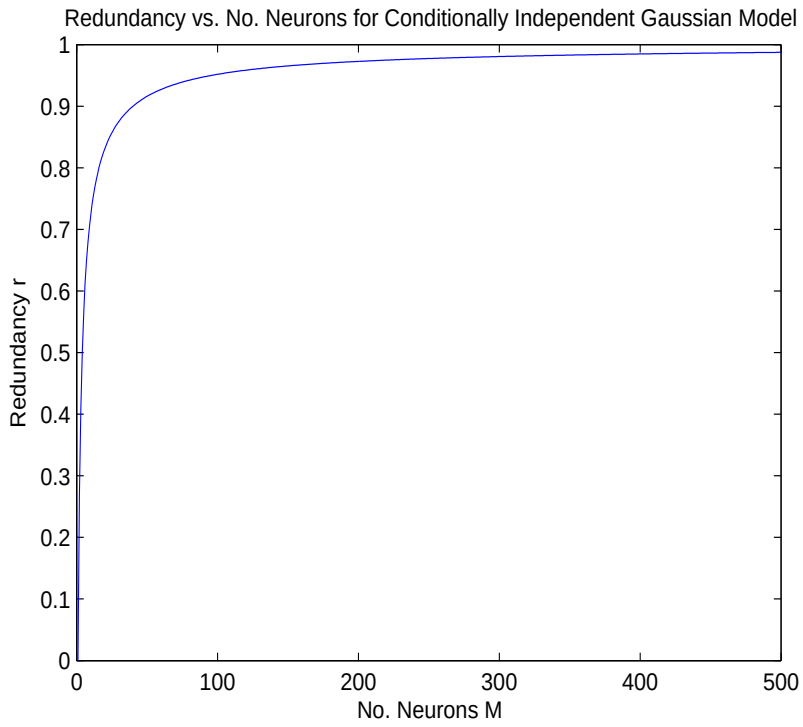


Figure 2.2. redundancy vs. number of neurons for conditionally independent gaussian model

Definition. *The function* $f : \mathbb{R}^2 \to \mathbb{R}$. *The stimulus* S *is a random variable. The noise* W_i *are independent random variables and independent of* S. *Each response* $R_i = f(S, W_i)$. *It follows that the population response* $R_1, R_2, ..., R_M$ *are identically distributed random variables which are conditionally independent given* S. *Figure 2.3 illustrates this model.*

Property 5. *Let* r_M *denote the population redundancy of* M *neurons. Let* $\delta_M = r_{M+1} - r_M$ *be the difference in redundancy when an additional neuron is added to the population. For the general conditional independent model, for* $M \in \mathbb{N}$

(i) r_M *is increasing in* M

(ii) $\lim_{M\to\infty} \delta_M = 0$

(iii) $r_M \leq 1$ *for all* M

Proof. (i) The goal is to show that $\delta_M \geq 0$. By definition, it follows that

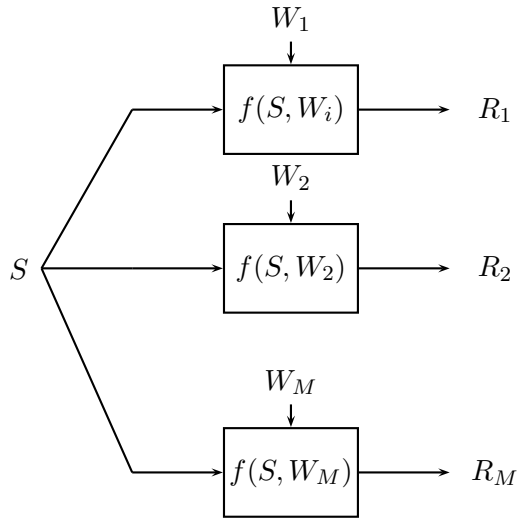


Figure 2.3. General Conditional Independent Model.

$$\delta_M = \frac{\sum_{i=1}^{M+1} H(R_i) - H(R_1, ..., R_{M+1})}{\sum_{i=1}^{M+1} I(S; R_i)} - \frac{\sum_{i=1}^{M} H(R_i) - H(R_1, ..., R_M)}{\sum_{i=1}^{M} I(S; R_i)} \tag{2.6}$$

For this model, it can be shown that

$$\sum_{i=1}^{M} H(R_i|R_{i-1}, ..., R_1) - MH(R_{M+1}|R_M, ..., R_1) \geq 0 \tag{2.7}$$

Therefore, it follows that $\delta_M \geq 0$.

(ii)

$$\delta_M = \frac{\sum_{i=1}^{M} H(R_i|R_{i-1}, ..., R_1) - H(R_{M+1}|R_M, ..., R_1)}{M(M+1)I(S; R_i)} \tag{2.8}$$

$$\leq \frac{\sum_{i=1}^{M} H(R_i|R_{i-1}, ..., R_1)}{M(M+1)I(S; R_i)} \tag{2.9}$$

$$\leq \frac{\sum_{i=1}^{M} H(R_i)}{M(M+1)I(S; R_i)}$$

$$= \frac{MH(R_i)}{M(M+1)I(S; R_i)}$$

The second inequality follows from the fact that condition

$$\lim_{M\to\infty} \delta_M \leq \lim_{M\to\infty} \frac{MH(R_i)}{M(M+1)I(S;R_i)} \tag{2.10}$$

$$\leq 0 \tag{2.11}$$

(iii) By definition,

$$r = \frac{\sum_{i=1}^{M} H(R_i) - H(R_1, ..., R_M)}{\sum_{i=1}^{M} I(S;R_i)} \tag{2.12}$$

$$\tag{2.13}$$

The fact that condition

$$H(R_1, ..., R_M) \leq H(R_1, ..., R_M|S). \tag{2.14}$$

Therefore,

$$r \leq \frac{\sum_{i=1}^{M} H(R_i) - H(R_1, ..., R_M|S)}{\sum_{i=1}^{M} I(S;R_i)} \tag{2.15}$$

$$\tag{2.16}$$

The chain rule of mutual information combined with the fact that conditioning reduces entropy implies that

$$H(R_1, ..., R_M|S) = \sum_{i}^{M} H(R_i|S) \tag{2.17}$$

Using this simplification,

$$r \leq \frac{\sum_{i=1}^{M} H(R_i) - \sum_{i}^{M} H(R_i|S)}{\sum_{i=1}^{M} I(S;R_i)} \tag{2.18}$$

$$= 1 \tag{2.19}$$

□

2.3 Multiple Clusters Gaussian Model

In the previous two examples, the stimulus is modeled by a single variable. However, in actual settings, the stimulus is often multidimensional, and different responses encode

distinct facets of the stimulus. For example, the spectrogram (spatio-temporal representation) of an audio wave stimulus is multidimensional. Neurons from different regions of the auditory system may be sensitive to different spatio-temporal aspects of the stimulus. This model captures the population behavior to multidimensional stimulus by modeling the stimulus as a random vector and by separating the responses into clusters that react to different aspects of the stimulus.

Notation:

- n: number of neural clusters
- $\mathbf{S}$: vector stimulus $(S_1, ..., S_n)$
- M_i: number of neurons in cluster i
- $W_j^{(i)}$: noise associated with neuron j in cluster i
- $R_j^{(i)}$: response of neuron j in cluster i
- $\mathbf{W^{(i)}}$: noise for cluster i $\left(W_1^{(i)}, W_2^{(i)}, ..., W_{M_i}^{(i)}\right)$
- $\mathbf{R^{(i)}}$: response of cluster i $\left(R_1^{(i)}, R_2^{(i)}, ..., R_{M_i}^{(i)}\right)$

Definition. *In a neural population with n clusters, for each $i \in \{1, ..., n\}$, cluster i contains M_i neurons. The total number of neurons $M = M_1 + M_2 + ... + M_n$. The stimulus $\mathbf{S} = (S_1, S_2, ..., S_n)$ where S_i are i.i.d $\mathcal{N}(0, \sigma_s^2)$. The noise $W_j^{(i)}$ are i.i.d $\sim \mathcal{N}(0, \sigma^2)$ and independent of S. For each cluster i, let the responses $R_j^{(i)} = S_i + W_j^{(i)}$. Therefore, for each cluster, the responses are gaussian random variables conditionally independent given S. In addition, the responses of different clusters are independent. Figure 2.4 illustrates this model.*

Since all the responses $R_j^{(i)}$ are gaussian with variance $\sigma_s^2 + \sigma^2$, the entropy of a response is defined as

$$H(R_j^{(i)}) = \frac{1}{2} \log(2\pi e)(\sigma_s^2 + \sigma^2) \tag{2.20}$$

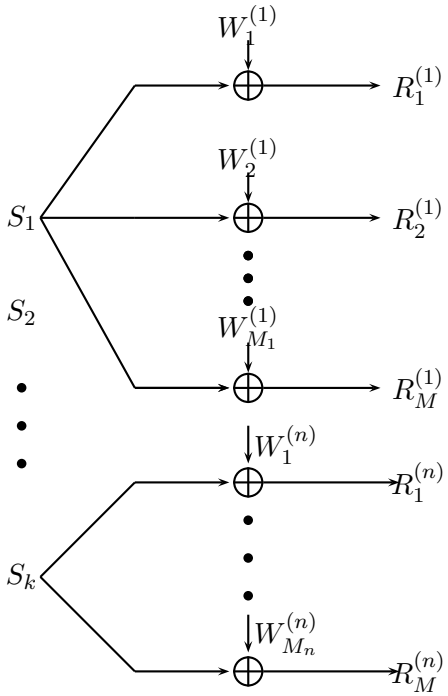


Figure 2.4. Multiple Clusters Gaussian Model.

Since the the responses of the n clusters $\mathbf{R^{(1)}}, \mathbf{R^{(2)}}, ..., \mathbf{R^{(n)}}$ are independent, the joint entropy is given by

$$\begin{aligned} H(\mathbf{R^{(1)}}, ..., \mathbf{R^{(n)}}) &= \sum_{i=1}^{n} H(\mathbf{R^{(n)}}) \\ &= \frac{M}{2}\log(2\pi e) + \frac{M-n}{2}\log\sigma^2 + \sum_{i=1}^{n}\frac{1}{2}\log(M_i\sigma_s^2 + \sigma^2) \end{aligned} \tag{2.21}$$

The mutual information between a S and $R_j^{(i)}$ is given by

$$I(S; R_j^{(i)}) = \frac{1}{2}\log(1 + \frac{\sigma_s^2}{\sigma^2}) \tag{2.22}$$

Using the above three measures, the population redundancy r can be found

$$r = 1 - \frac{\sum_{i=1}^{n}\log(\frac{M_i\sigma_s^2}{\sigma^2} + 1)}{M\log(1 + \frac{\sigma_s^2}{\sigma^2})} \tag{2.23}$$

Property 6. *For the Multiple Clusters Model,*

(i) $0 \leq r \leq 1$

(ii) $\lim_{M\to\infty} r = 1$

Proof. (i) By construction, the responses are conditionally independent given the stimulus. Therefore, the proof is same as that for the conditionally independent gaussian model.
(ii) For large M, the behavior of r becomes

$$r \sim 1 - \frac{n\log M}{M} \tag{2.24}$$

Therefore, $r \to 1$ as $M \to \infty$. □

From the properties of r, it can be seen that the redundancy of multiple clusters gaussian model behaves similar to the conditionally independent gaussian model.

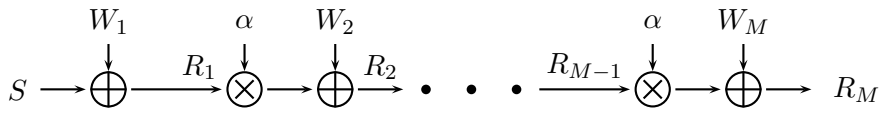


Figure 2.5. Gauss Markov Model.

2.4 Gauss Markov Model

In all the models considered so far, the responses are conditionally independent given the stimulus. The population redundancy r was seen to behave nicely and be bounded by one for all the models. This example examines the redundancy when the responses remain dependent given the stimulus by modeling the responses as a gauss-markov process.

Definition. *The stimulus* $S \sim \mathcal{N}(0,1)$. $0 \leq \alpha \leq 1$. *The noise* W_i *are i.i.d* $\mathcal{N}(0, 1-\alpha^2)$. *The response of the first neuron* $R_1 = S + W_1$. *For neuron* i *where* $i \geq 2$, *the response* $R_i = \alpha R_{i-1} + W_i$. *From construction, the responses* $R_1, R_2, ..., R_M$ *form a gauss markov process. In addition,* $R_1, R_2, ..., R_M$ *are identically distributed (but not independent)* $\mathcal{N}(0,1)$. *Figure 2.5 provides an illustration.*

Since for all i, $R_i \sim \mathcal{N}(0,1)$. The entropy of of a single response is given as

$$H(R_i) = \frac{1}{2}\log(2\pi e) \tag{2.25}$$

Using the chain rule of entropy, the joint entropy of $R_1, R_2, ..., R_M$ can be rewritten as

$$H(R_1, ..., R_M) = \sum_{i=1}^{M} H(R_i|R_{i-1}, ..., R_1) \tag{2.26}$$

$$\tag{2.27}$$

Since $R_1, R_2, ..., R_M$ form a markov process, R_i is conditionally independent of $R_{i-2}, ..., R_1$ given R_{i-1}. Therefore,

$$\sum_{i=1}^{M} H(R_i|R_{i-1}, ..., R_1) = \sum_{i=1}^{M} H(R_i|R_{i-1}) \tag{2.28}$$

Since by construction $R_i = R_{i-1} + W_i$ for all $i > 1$, it follows that $H(R_i|R_{i-1}) = H(W_i)$. Also, since W_i are i.i.d $\sim \mathcal{N}(0, 1-\alpha^2)$. It follows that

$$\sum_{i=1}^{M} H(R_i|R_{i-1}) = H(R_1) + (M-1)H(W_i) \tag{2.29}$$

$$= \frac{1}{2}\log \pi e + (M-1)\frac{1}{2}\log \pi e(1-\alpha^2) \tag{2.30}$$

The mutual information between S and R_i is defined as

$$I(S; R_i) = H(R_i) - H(R_i|S) \tag{2.31}$$

Using recursion, R_i can be rewritten as

$$R_i = \alpha^i S + \sum_{j=0}^{i-1} \alpha^j W_{i-j} \tag{2.32}$$

Substituting equation 2.32 into 2.31, the mutual information becomes

$$I(S; R_i) = H(R_i) - H(\alpha^i S + \sum_{j=0}^{i-1} \alpha^j W_{i-j}|S) \tag{2.33}$$

$$= H(R_i) - H(\sum_{j=0}^{i-1} \alpha^j W_{i-j}|S)$$

Using the fact that W_i are i.i.d $\mathcal{N}(0, 1-\alpha^2)$ and independent of S,

$$\begin{aligned} I(S;R_i) &= H(R_i) - H(\sum_{j=0}^{i-1} \alpha^j W_{i-j}) \\ &= -\frac{1}{2}\log(1-\alpha^{2i}) \end{aligned} \tag{2.34}$$

Using the above information theory quantities, the population redundancy r is found to be

$$r = \frac{\frac{M}{2}\log 2\pi e - \frac{1}{2}\log 2\pi e + (M-1)\frac{1}{2}\log 2\pi e(1-\alpha^2)}{\sum_{i=1}^{M}\frac{1}{2}\log(\frac{1}{1-\alpha^{2i}})} \tag{2.35}$$

Property 7. *For the gauss markov model,* $\lim_{M\to\infty} r = \infty$

Proof. It can be shown that $\sum_{i=1}^{\infty}\log(\frac{1}{1-\beta^i})$ converges. Therefore, for large M,

$$r \sim M \tag{2.36}$$

□

Figure 2.6 confirms the result that for large M, r scales linearly with M. Unlike the conditionally independent model, the population redundancy r for the gauss markov model is not restricted lie within the $[0, 1]$ interval. In fact, r can be unbounded. The gauss markov model portrays an instance when the redundancy measure from definition 16 fails to lie within $[0, 1]$, the desired interval.

2.5 Two Layers Mixture Model

Unlike the conditionally independent gaussian model, the gauss markov model does not give a nice behavior of redundancy. This model combines the conditionally independent gaussian model and the gauss markov model by separating the responses into two layers. The first layer has a conditionally independent gaussian structure while the second layer has a gauss markov structure. This model is consistent with the intuition that actual neural systems contains multiple layers that perform joint processing of information.

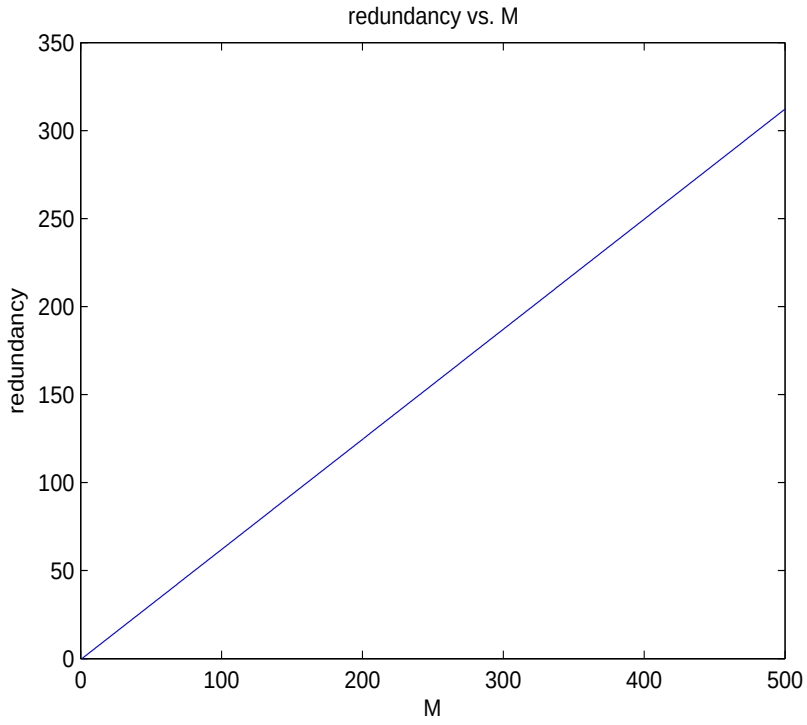


Figure 2.6. Redundancy vs. No. Neurons for Gauss Markov Model

Notation

- M_i: number of neurons in cluster i
- W_j^i: noise associated with neuron j in cluster i
- R_j^i: response of neuron j in cluster i
- $\mathbf{W^i}$: noise for cluster i $\left(W_1^i, W_2^i, ..., W_{M_i}^i\right)$
- $\mathbf{R^i}$: response of cluster i $\left(R_1^i, R_2^i, ..., R_{M_i}\right)$
- $\mathbf{A}$: $M_1 \times M_2$ matrix with each row having the same number of non zeros elements
- k: sparsity of each row of A

Definition. *The sparsity of a vector* $\mathbf{X}$ *is the number of non zero elements in* $\mathbf{X}$.

Definition. *The neural population contains two layers with size* $M_1 \times M_2$ *respectively. The total number of neurons* $M = M_1 + M_2$. *The stimulus* $S \sim \mathcal{N}(0, \alpha^2)$ *for* $\alpha \leq 1$.

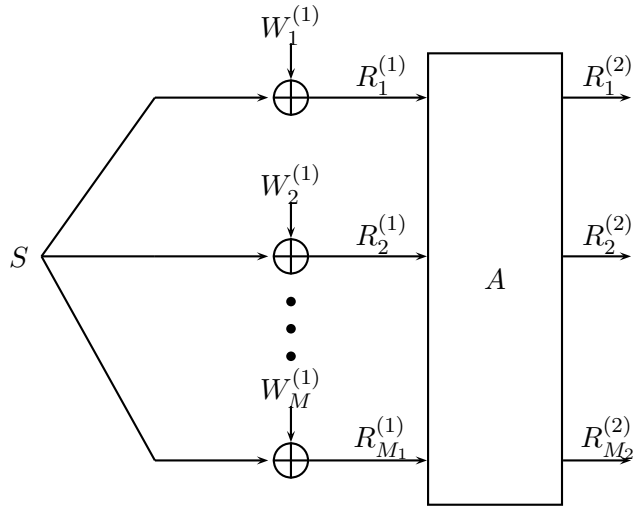


Figure 2.7. Two Layer Mixture Model.

The noises of the first layer $W_1^1, ..., W_{M_1}^1$ *are i.i.d* $\mathcal{N}(0, 1-\alpha^2)$*. Each response of the first layer* $R_i^1 = S_i + W_i^1$*. It follows that the responses of the first layer are gaussian random variables conditionally independent given* S*. The noises in the second layer* $W_1^2, ..., W_{M_2}^2$ *are i.i.d* $\mathcal{N}(0, 1-\alpha^2)$ *and independent of noises of the first layer* $\mathbf{W^1}$*. Consider* $\mathbf{A}$ *to be a* $M_1 \times M_2$ *matrix with each row having the same sparsity* k*. That is, each row of* A *has* k *non zero elements. Each non zero element* $A_{ij} = \frac{1}{k}$*. The responses of the second layer,* $\mathbf{R^2} = \mathbf{AR^1} + \mathbf{W^2}$*. Thus, each response* R_i^2 *is formed by linearly combining* k *responses from the first layer and corrupting with noise. Figure 2.7 gives an illustration.*

As given in model definition, each neuron in layer one $R_i^{(1)}$ can be represented as

$$R_i^{(1)} = S_i + W_i^{(1)} \tag{2.37}$$

and each neuron in layer two $R_i^{(2)}$ can be represented as

$$\begin{aligned} R_i^{(2)} &= \sum_{j=1}^{M_1} A_{ij} R_j + Z_i \\ &= S \sum_{j=1}^{M_1} A_{ij} + \sum_{j=1}^{M_1} A_{ij} W_j + Z_i \end{aligned} \tag{2.38}$$

The covariances of $R_i^{(1)}$ and $R_i^{(2)}$ are

$$Cov[R_i^{(1)}] = 1 \tag{2.39}$$

$$Cov[R_j^{(2)}] = 1 + \frac{(1-\alpha^2)}{k} \tag{2.40}$$

The individual entropies of $R_i^{(1)}$ and $R_i^{(2)}$ can be evaluated

$$H(R_i^{(1)}) = \frac{1}{2}\log 2\pi e \tag{2.41}$$

$$H(R_i^{(2)}) = \frac{1}{2}\log 2\pi e(1 + \frac{1-\alpha^2}{k}) \tag{2.42}$$

The joint entropy of the population response $R_1^{(1)}, ..., R_{M_1}^{(1)}, R_1^{(2)}, ..., R_{M_2}^{(2)}$ is found to be

$$H(R_1, ..., R_{M_1}, R_1^{(2)}, ..., R_{M_2}^{(2)}) = H(R_1, ..., R_{M_1}) + H(R_1^{(2)}, ..., R_{M_2}^{(2)}|R_1, ..., R_{M_1}) \tag{2.43}$$

$$= \frac{1}{2}\log(2\pi e)^{M_1}(M_1\alpha^2 + (1-\alpha^2))(1-\alpha^2)^{M_1-1} \tag{2.44}$$

$$+ \frac{M_2}{2}\log 2\pi e(1-\alpha^2) \tag{2.45}$$

$$= \frac{M_1+M_2}{2}\log(2\pi e) + \frac{M_1+M_2}{2}\log(1-\alpha^2)$$

$$+ \frac{1}{2}\log(M_1\alpha^2 + (1-\alpha^2)) - \frac{1}{2}\log(1-\alpha^2) \tag{2.46}$$

The signal-to-noise ratio of responses in layer one SNR_1 and layer two SNR_2 are respectively

$$SNR_1 = \frac{\alpha^2}{1-\alpha^2} \tag{2.47}$$

$$SNR_2 = \frac{\alpha^2(\sum_{j=1}^{M_1} A_{ij})^2}{(1-\alpha^2)\sum_{j=1}^{M_1} A_{ij}^2 + (1-\alpha^2)} \tag{2.48}$$

$$= \frac{\alpha^2}{\frac{1}{k}(1-\alpha^2) + (1-\alpha^2)} \tag{2.49}$$

Therefore, the mutual information between S and a single response in layer one $R_i^{(1)}$ is

$$I(S; R_i^{(1)}) = \frac{1}{2}\log(1 + \frac{\alpha^2}{1-\alpha^2}) \tag{2.50}$$

and the mutual information between S and a single response in layer two $R_i^{(2)}$ is

$$I(S; R_i^{(2)}) = \frac{1}{2}\log(1 + \frac{\alpha^2}{\frac{1-\alpha^2}{k} + 1 - \alpha^2}) \tag{2.51}$$

Using the measure evaluated above, the population redundancy r is found to be

$$r = \frac{\frac{M_2}{2}\log(1+\frac{(1-\alpha^2)}{k}) + \frac{M_1+M_2}{2}\log(\frac{1}{1-\alpha^2}) - \frac{1}{2}\log(M_1\alpha^2+(1-\alpha^2)) - \frac{1}{2}\log(\frac{1}{1-\alpha^2})}{\frac{M_1}{2}\log(1+SNR_1)+\frac{M_2}{2}\log(1+SNR_2)} \tag{2.52}$$

$$= \frac{\frac{M_2}{2}\log(1+\frac{(1-\alpha^2)}{k}) + \frac{M_1+M_2}{2}\log(\frac{1}{1-\alpha^2}) - \frac{1}{2}\log(M_1\alpha^2+(1-\alpha^2)) - \frac{1}{2}\log(\frac{1}{1-\alpha^2})}{\frac{M_1}{2}\log(1+\frac{\alpha^2}{1-\alpha^2})+\frac{M_2}{2}\frac{1}{2}\log(1+\frac{\alpha^2}{\frac{1-\alpha^2}{k}+1-\alpha^2})} \tag{2.53}$$

Property 8. *For the two layer mixture model, for $k \sim M_1$, $M_1, M_2 \sim M$,*

$$lim_{M\to\infty} r = 1 \tag{2.54}$$

Proof. Since $k \sim M_1$ and $M_1 \sim M$,

$$\lim_{M\to\infty} SNR_2 = \frac{\alpha^2}{1-\alpha^2} \tag{2.55}$$

Substituting this into 2.52 and simplifying it can be seen that

$$r \sim \frac{\frac{M_2}{2}\log(1+\frac{(1-\alpha^2)}{k}) + \frac{M}{2}\log(\frac{1}{1-\alpha^2})}{\frac{M}{2}\log(\frac{1}{1-\alpha^2})} \tag{2.56}$$

Since $k \sim M_1$, it follows that

$$r \to 1 \tag{2.57}$$

□

Figures 2.9 and 2.8 show plots of r versus M for the case where $M_1 = M_2 = \frac{1}{2}M$. The figures show that for $M \leq 30$, $r \geq 1$, and for $M \geq 30$, $r \leq 1$. It appears that for small values of M, the mixture model behaves like the gauss markov model (discussed in section 2.4), and for large values of M, the mixture model behaves like the conditionally independent gaussian model (discussed in section 2.1).

2.A Proof Details

This section presents some of the details of proofs in this chapter.

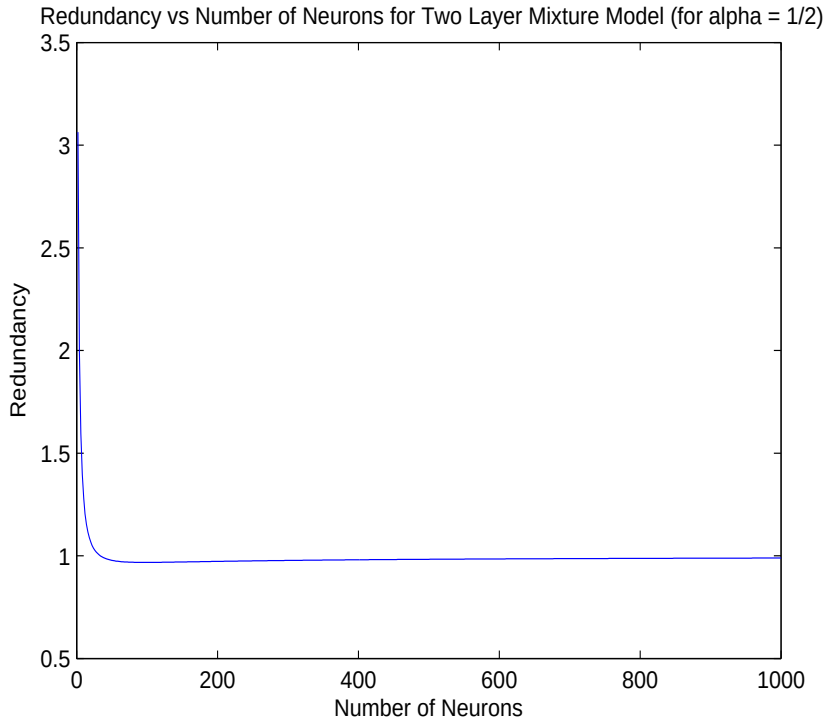


Figure 2.8. Redundancy vs. No. Neurons for the Two Layer Mixture Model

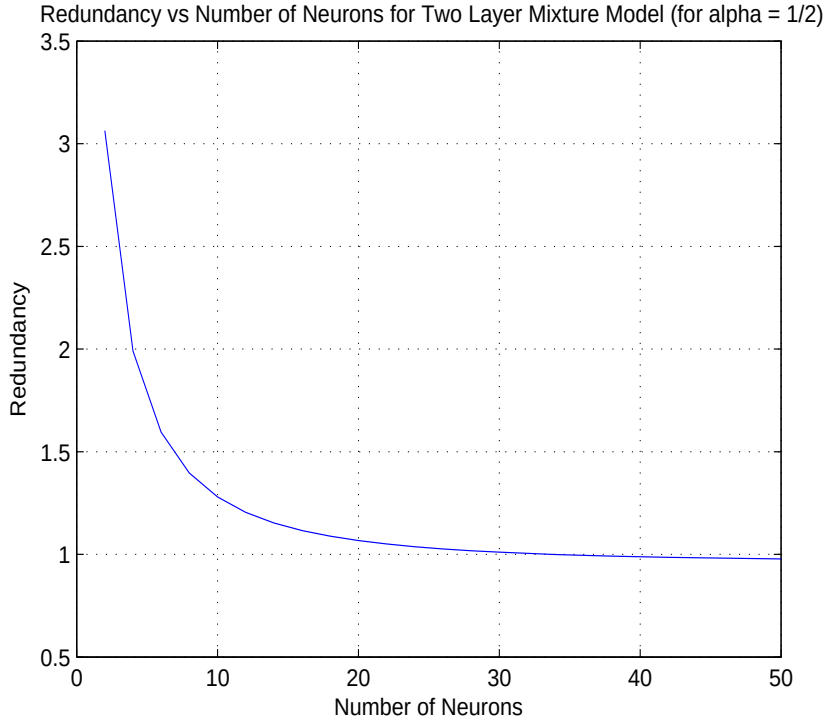


Figure 2.9. Redundancy vs. No. Neurons for the Two Layer Mixture Model

2.5.1 Conditionally Independent Gaussian Model

I. Redundancy Measure Evaluation

$$r = \frac{\frac{M}{2}\log 2\pi e(\sigma_s^2+\sigma_w^2) - (\frac{M}{2}\log 2\pi e + \frac{1}{2}(M\sigma_s^2+\sigma_w^2) + \frac{M-1}{2}\log\sigma_w^2)}{\frac{M}{2}\log(1+\frac{\sigma_s^2}{\sigma_w^2})} \tag{2.58}$$

$$= \frac{\frac{M}{2}\log(\sigma_s^2+\sigma_w^2) - \frac{1}{2}\log(M\sigma_s^2+\sigma_w^2) - \frac{M-1}{2}\log\sigma_w^2}{\frac{M}{2}\log(1+\frac{\sigma_s^2}{\sigma_w^2})} \tag{2.59}$$

$$= \frac{\frac{M}{2}\log(1+\frac{\sigma_s^2}{\sigma_w^2}) - \frac{1}{2}\log(1+\frac{M\sigma_s^2}{\sigma_w^2})}{\frac{M}{2}\log(1+\frac{\sigma_s^2}{\sigma_w^2})}$$

$$= 1 - \frac{\log(1+\frac{M\sigma_s^2}{\sigma_w^2})}{M\log(1+\frac{\sigma_s^2}{\sigma_w^2})}$$

2.5.2 General Conditional Independent Model

I. δ_M Evaluation

$$\delta_M = \frac{(M+1)H(R_i) - H(R_1,...,R_{M+1})}{(M+1)I(S;R_i)} - \frac{MH(R_i) - H(R_1,...,R_M)}{MI(S;R_i)} \tag{2.60}$$

$$= \frac{(M+1)H(R_i) - H(R_1,...,R_M) - H(R_{M+1}|R_1,...,R_M)}{(M+1)I(S;R_i)} - \frac{MH(R_i) - H(R_1,...,R_M)}{MI(S;R_i)} \tag{2.61}$$

$$= \frac{H(R_1,...,R_M)}{I(S;R_i)}(\frac{1}{M} - \frac{1}{M+1}) - \frac{H(R_{M+1}|R_M,...,R_1)}{(M+1)I(S;R_i)}$$

$$= \frac{H(R_1,...,R_M) - MH(R_{M+1}|R_M,...,R_1)}{M(M+1)I(S;R_i)} \tag{2.62}$$

$$= \frac{\sum_{i=1}^{M} H(R_i|R_{i-1},...,R_1) - MH(R_{M+1}|R_M,...,R_1)}{M(M+1)I(S;R_i)}$$

II. Proof for Property 5(i)

Since $R_1, R_2, ..., R_M, R_{M+1}$ are identically distributed,

$$H(R_i|R_{i-1},...,R_1) = H(R_{M+1}|R_{i-1},...,R_1) \tag{2.63}$$

Since conditiong reduces entropy, for all $i \leq M+1$,

$$H(R_{M+1}|R_M, ..., R_1) \leq H(R_{M+1}|R_{i-1}, ..., R_1) \tag{2.64}$$

Combining equations 2.63 and 2.64, and substituting into equation 2.6, it can be found that

$$\sum_{i=1}^{M} H(R_i|R_{i-1}, ..., R_1) - MH(R_{M+1}|R_M, ..., R_1) \geq 0 \tag{2.65}$$

2.5.3 multiple clusters gaussian model

I. Redundancy Measure Evaluation

$$r = \frac{\frac{M}{2}\log(2\pi e)(\sigma_s^2+\sigma^2) - \frac{M}{2}\log(2\pi e) - \frac{M-n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{1}{2}\log(M_i\sigma_s^2+\sigma^2)}{\frac{M}{2}\log(1+\frac{\sigma_s^2}{\sigma^2})} \tag{2.66}$$

$$= \frac{\frac{M}{2}\log(1+\frac{\sigma_s^2}{\sigma^2}) + \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\log(M_i\sigma_i^2+\sigma^2)}{\frac{M}{2}\log(1+\frac{\sigma_s^2}{\sigma^2})} \tag{2.67}$$

$$= 1 - \frac{\sum_{i=1}^{n}\log(\frac{M_i\sigma_s^2}{\sigma^2}+1)}{M\log(1+\frac{\sigma_s^2}{\sigma^2})}$$

II. Proof for Property 6

Let $\beta = \alpha^2$. The goal is to show the convergence of the term $\sum_{i=1}^{\infty}\log(\frac{1}{1-\beta^i})$. Using the fact that $\log x \leq x-1$, it follows that

$$\sum_{i=1}^{\infty}\log\left(\frac{1}{1-\beta^i}\right) \leq \sum_{i=1}^{\infty}\left(\frac{1}{1-\beta^i}-1\right) \tag{2.68}$$

$$= \sum_{i=1}^{\infty}\frac{\beta^i}{1-\beta^i} \tag{2.69}$$

The series $\sum_{i=1}^{\infty}\frac{\beta^i}{1-\beta^i}$ converges. To see this, use the ratio test (insert reference)

$$\lim_{n\to\infty}\left|\frac{\frac{\beta^{n+1}}{1-\beta^{n+1}}}{\frac{\beta^{n}}{1-\beta^{n}}}\right| = \lim_{n\to\infty}\left|\frac{\beta-\beta^{n+1}}{1-\beta^{n+1}}\right| \tag{2.70}$$

$$= \beta \tag{2.71}$$

$$\leq 1 \tag{2.72}$$

Since the term $\sum_{i=1}^{\infty}\log\left(\frac{1}{1-\beta^{i}}\right)$ converges, for large M,

$$r \sim M \tag{2.73}$$

Chapter 3

Joint Entropy Approximation

For most of the simple models presented in the previous chapter, precise calculation of the joint entropy of the population response $H(R_1, R_2, ..., R_M)$ is be feasible. However, in neural systems, an accurate estimate of the joint distribution of the population response $p(R_1, R_2, ..., R_M)$ is necessary in order to calculate the joint entropy. However, for large M, $p(R_1, R_2, ..., R_M)$ is a high dimensional probability distribution. Therefore, an accurate estimate is infeasible due to limited neural data. Therefore, methods are needed to approximate the joint entropy. In this chapter, a few different approximations for the joint entropy of the population response are presented. In addition, the performance of the approximations are analyzed for the conditionally independent gaussian model defined in section 2.1 and the gauss markov model defined in section 2.4. These approximations contain only pairwise entropies $H(R_i, R_j)$ and single entropies $H(R_i)$. These low dimensional entropies can be calculated more accurately since only pairwise distributions $p(R_i, R_j)$ and single distributions $p(R_i)$ are needed.

3.1 Approximation 1

Definition. *The first approximation of the joint entropy of the population response* $\hat{H}_1(\mathbf{R})$ *is defined as*

$$\hat{H}_1(\mathbf{R}) = \frac{2}{M-1}\sum_{i>j}^{M} H(R_i, R_j) - \sum_{i=1}^{M} H(R_i) \tag{3.1}$$

Property 9. *The first approximation* $\hat{H}_1(\mathbf{R})$ *is perfect if* $R_1, R_2, ..., R_M$ *are independent.*

Proof. If $R_1, R_2, ..., R_M$ are independent, then for all i, j,

$$H(R_i, R_j) = H(R_i) + H(R_j) \tag{3.2}$$

Making this substitution in the definition of $\hat{H}_1(\mathbf{R})$ and simplifying, it follows that

$$\hat{H}_1(\mathbf{R}) = \sum_{i=1}^{M} H(R_i) \tag{3.3}$$

□

Property 10. *For* $M \geq 3$*, for all joint distributions* $p(\mathbf{R})$*,* $\hat{H}_1(\mathbf{R}) \geq 0$*.*

Proof. Since, for all i, j, $H(R_i, R_j) \geq \max\{H(R_i), H(R_j)\}$, it follows that

$$\frac{2}{M-1}\sum_{i>j}^{M} H(R_i, R_j) \geq \sum_{i=1}^{M} H(R_i) \tag{3.4}$$

□

For the conditionally independent gaussian model defined in section 2.1, the estimated joint entropy $\hat{H}_1(\mathbf{R})$ evaluates to

$$\hat{H}_1(\mathbf{R}) = \frac{M}{2}\log(2\pi e) + \frac{M}{2}\log\left(\frac{2\sigma^2 + \sigma_w^2}{\sigma^2 + \sigma_w^2}\right) + \frac{M}{2}\log(\sigma_w^2) \tag{3.5}$$

The estimated joint entropy will be compared to the actual joint entropy of the conditionally independent gaussian model. The performance of the approximation will be judged based on the error-ratio.

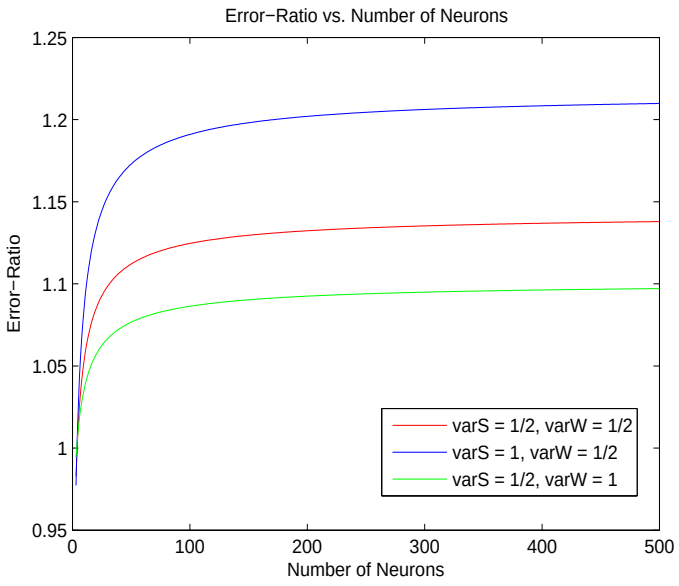


Figure 3.1. Error-ratio vs. M for conditionally independent gaussian model for approximation 1

Definition 18. *Given random variables $R_1, R_2, ..., R_M$, the error-ratio E_r of an joint entropy approximation $\hat{H}(R_1, R_2, ..., R_M)$ is defined to be*

$$E_r = \frac{\hat{H}(R_1, R_2, ..., R_M)}{H(R_1, R_2, ..., R_M)} \tag{3.6}$$

Note that if the estimator is perfect, then $E_r = 1$.

Figure 3.1 plots E_r vs. M for different values of σ_s^2 and σ_w^2. The result shows that estimator performs well since E_r approaches a value slightly above one as for large M.

Figure 3.2 plots E_r vs. σ_w^2 for fixed M and σ_s^2. The figure shows that $\hat{H}_1$ is a good estimator since E_r approaches a value slightly above one for large M.

Figure 3.3 plots E_r vs. σ_s^2 for fixed M and σ_w^2. Like the previous two figures, this figures shows that $\hat{H}_1$ performs well since E_r approaches a value slightly above one for large M.

Let $\mathbf{K_r}$ denote the covariance matrix of $\mathbf{R}$. For the gauss markov model defined in section 2.4, the estimated joint entropy $\hat{H}_1(\mathbf{R})$ evaluates to

$$\hat{H}_1(\mathbf{R}) = \frac{2}{M-1} \sum_{i=1}^{M-1} i \frac{1}{2} \log(2\pi e)^2 (1 - \alpha^{2(M-i)}) - \frac{M}{2} \log(2\pi e) \tag{3.7}$$

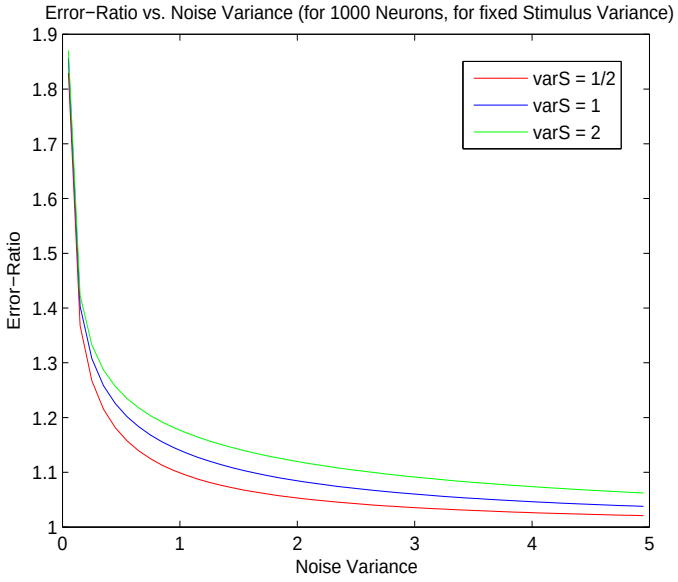


Figure 3.2. Error-ratio vs. σ_w^2 for conditionally independent gaussian model for approximation 1

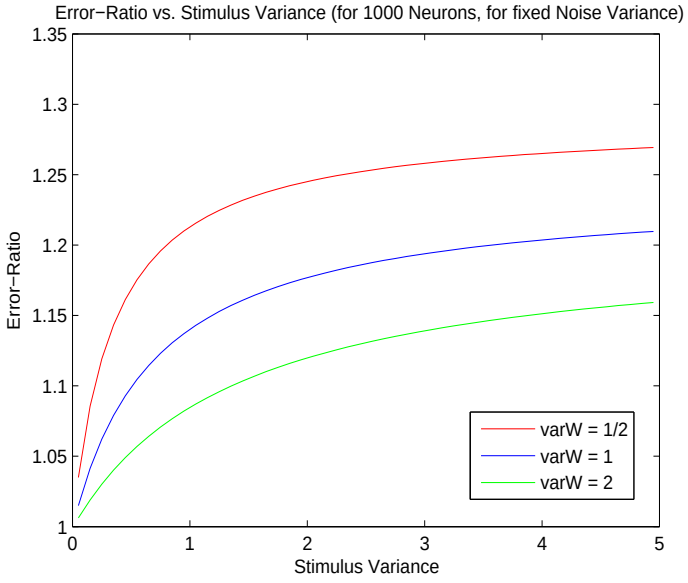


Figure 3.3. Error-ratio vs. σ^2 for conditionally independent gaussian model for approximation 1

$$= \frac{M}{2}\log(2\pi e) + \frac{1}{M-1}\sum_{i=1}^{M-1} i\log(1-\alpha^{2(M-i)}) \tag{3.8}$$

Figure 3.4 plots E_r vs. M for different values of α. This figures shows that $\hat{H}_1$ is a good estimator since E_r approaches a value slightly above one for large M.

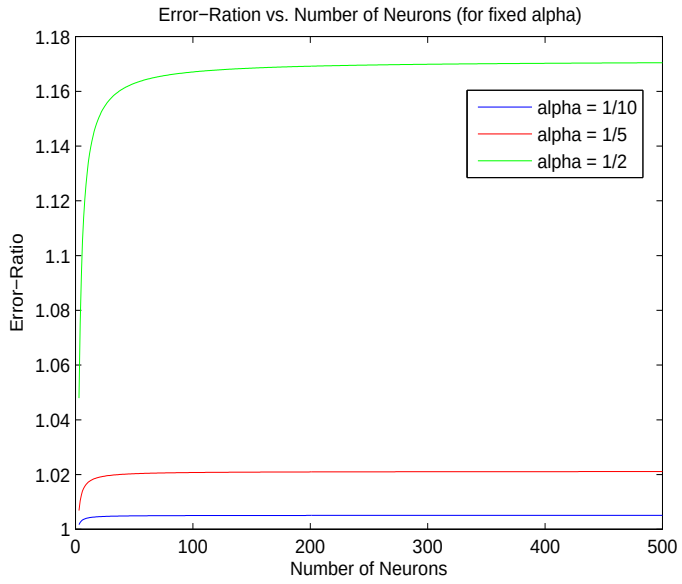


Figure 3.4. Error-ratio vs M for gauss markov model for approximation 1

Figure 3.5 plots E_r vs. α for fixed M. This figures shows that $\hat{H}_1$ is a good estimator for small values of α.

3.2 Approximation 2

Definition. *The second approximation of the joint entropy* $\hat{H}_2(\mathbf{R})$ *of the population response is defined as*

$$\hat{H}_2(\mathbf{R}) = \sum_{i=1}^{M} H(R_i) - \sum_{i>j} I(R_i; R_j) \tag{3.9}$$

Property 11. $\hat{H}_2(\mathbf{R}) = H(\mathbf{R})$ *if* $R_1, R_2, ..., R_M$ *are independent*

Proof. if $R_1, R_2, ..., R_M$ are independent, then for all i, j, $I(R_i, R_j) = 0$. Therefore,

$$\hat{H}_2(\mathbf{R}) = \sum_{i=1}^{M} H(R_i) \tag{3.10}$$

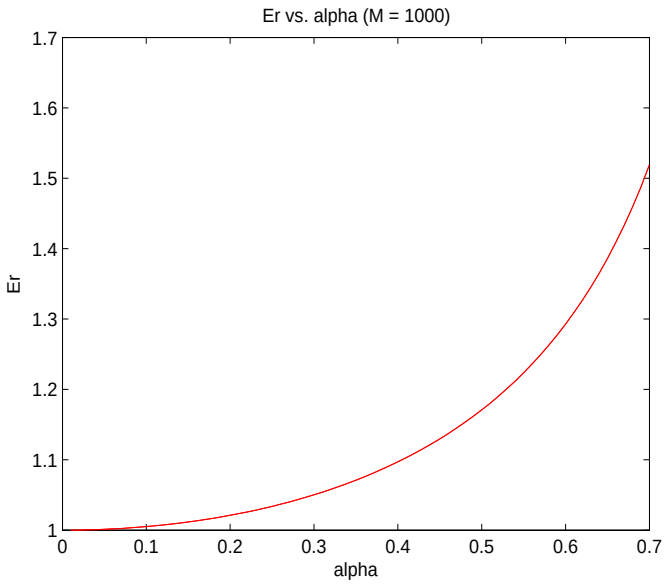


Figure 3.5. Error-ratio vs. α for gauss markov model for approximation 1

□

The following example shows that for certain distributions of $R_1, R_2, ..., R_M$, $\hat{H}_2(\mathbf{R})$ can be negative. Therefore, it is not a good estimator for those distributions.

Example 2. *Let $R_1 = R_2 = ... = R_M$. The the estimated joint entropy becomes*

$$\hat{H}_2(\mathbf{R}) = MH(R_1) - \frac{M(M-1)}{2}H(R) \tag{3.11}$$

Therefore, $\hat{H}_2(\mathbf{R}) < 0$ for $M > 3$.

For the conditionally independent gaussian model, $\hat{H}_2(\mathbf{R})$ evaluates to

$$\hat{H}_2(\mathbf{R}) = \frac{M}{2}\log(2\pi e) + \frac{M}{2}\log(\sigma^2 + \sigma_w^2) - \frac{M(M-1)}{2}\log\left(\frac{\sigma^2 + \sigma_w^2}{\sqrt{\sigma_w^2(2\sigma^2 + \sigma_w^2)}}\right) \tag{3.12}$$

Figure 3.6 plots of E_r vs. M. Since the plots show that $\hat{H}_2(\mathbf{R})$ underestimates and can be negative, it is not a good approximation for the conditionally independent gaussian model.

For the Gauss Markov Model, $\hat{H}_2(\mathbf{R})$ evaluates to be

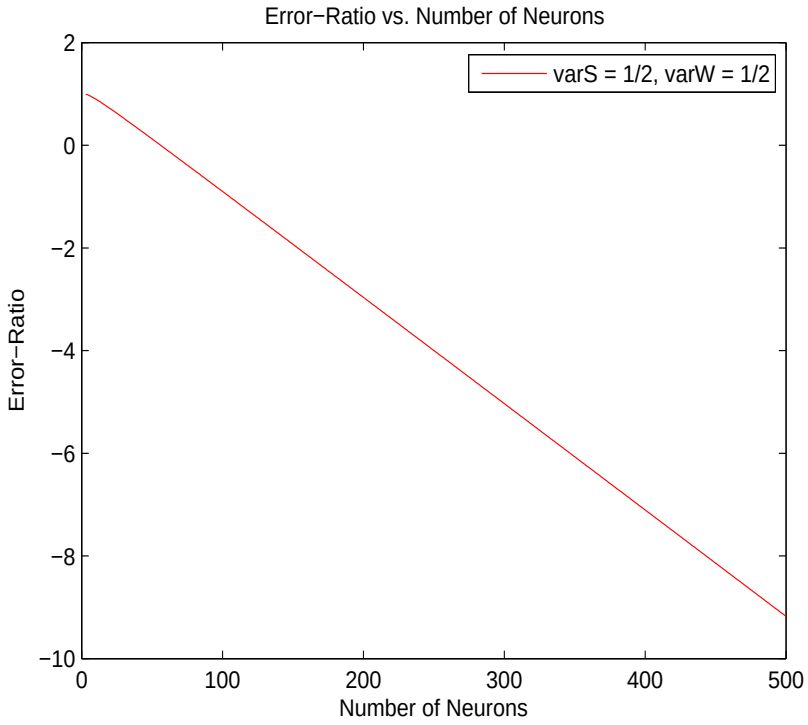


Figure 3.6. Error-ratio vs. M for conditionally independent gaussian model for approximation 2

$$\hat{H}_2(\mathbf{R}) = \frac{M}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^{M-1} i\log(1-\alpha^{2(M-i)}) \tag{3.13}$$

Figure 3.7 consists of plots of E_r vs. M for different values of α. Since E_r is close to one, $\hat{H}_2(R)$ performs well in estimating the entropy of a gauss markov process with $\alpha \leq 0.5$.

Figure 3.8 plots E_r vs. α for fixed M. the graph shows that for $\alpha \leq 0.7$, $\hat{H}_2(R)$ is a good entropy estimator of the gauss markov process.

3.3 Approximation 3

Definition. *The third approximation of the joint entropy* $\hat{H}_3(\mathbf{R})$ *of the population response is defined as*

$$\hat{H}_3(\mathbf{R}) = \sum_{i>j}^{M} H(R_i, R_j) - (M-2)\sum_{i=1}^{M} H(R_i) \tag{3.14}$$

Property 12. *If* $R_1, R_2, ..., R_M$ *are independent, then* $\hat{H}_3(\mathbf{R}) = H(\mathbf{R})$.

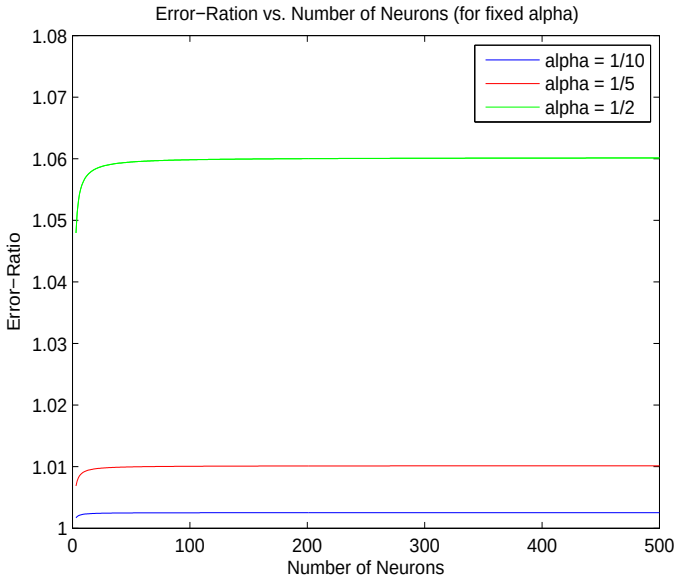


Figure 3.7. Error-ratio vs. M gauss markov model for approximation 2

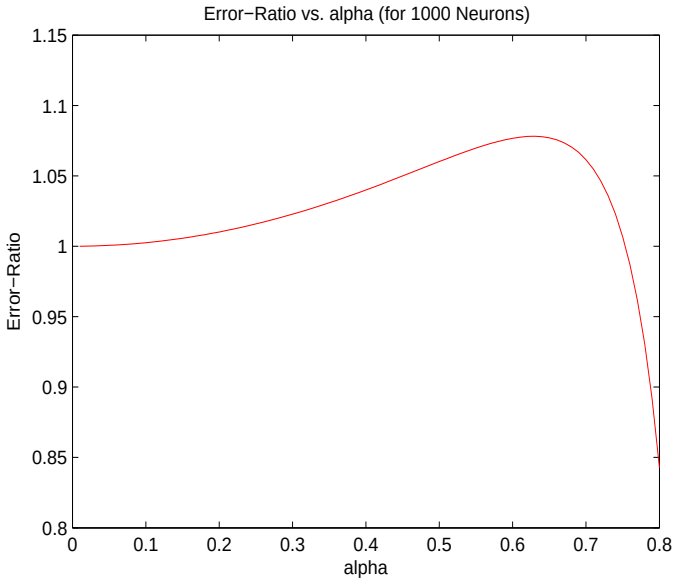


Figure 3.8. Error-ratio vs. α for gauss markov model for approximation 2

Proof. If $R_1, R_2, ..., R_M$, then for all i, j , $H(R_i, R_j) = H(R_i) + H(R_j)$. Therefore,

$$\hat{H}_3(\mathbf{R}) = (M-1)\sum_{i=1}^{M} H(R_i) - (M-2)\sum_{i=1}^{M} H(R_i) \tag{3.15}$$
$$= \sum_{i=1}^{M} H(R_i)$$

□

Similar to $\hat{H}_2(\mathbf{R})$, $\hat{H}_3(\mathbf{R})$ can be negative. Specifically, for model given in example 2, $\hat{H}_3(\mathbf{R}) < 0$ for $M > 3$.

For the conditionally independent gaussian model,$\hat{H}_3(\mathbf{R})$ evaluates to be

$$\hat{H}_3(\mathbf{R}) = \frac{M(M-1)}{4}\log((2\pi e)^2(2\sigma^2 + \sigma_w^2)\sigma_w^2) - \frac{M(M-2)}{2}\log(2\pi e(\sigma^2 + \sigma_w^2)) \tag{3.16}$$

Figure 3.9 plots E_r vs. M. It shows that $\hat{H}_3(\mathbf{R})$ underestimates and can be negative. Therefore, for the conditionally independent gaussian model, $\hat{H}_3(\mathbf{R})$ is a poor estimator of entropy.

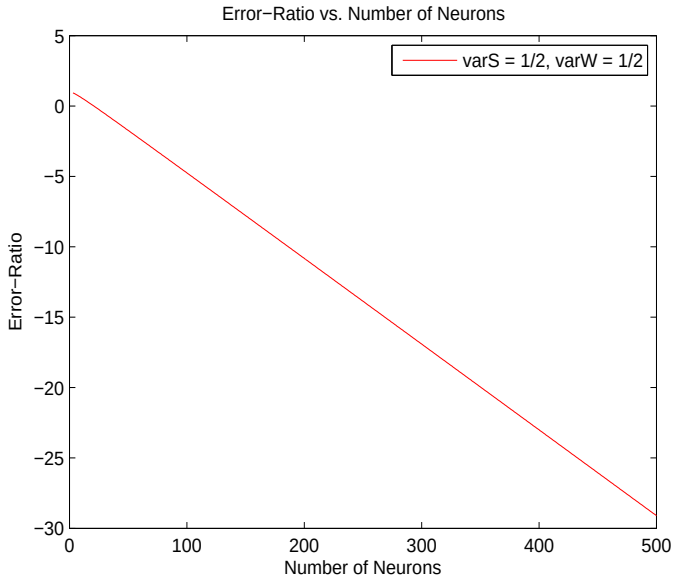


Figure 3.9. Error-ratio vs. M for conditionally independent gaussian model for approximation 3

For the gauss markov model, $\hat{H}_3(\mathbf{R})$ evaluates to be

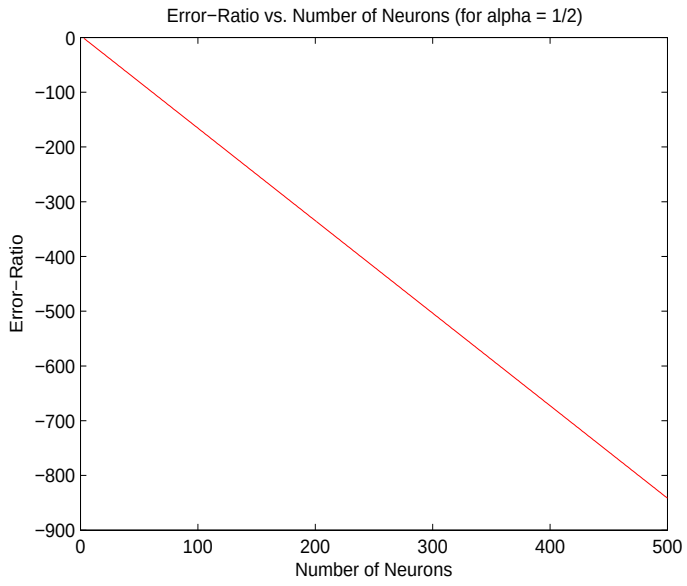


Figure 3.10. Error-ratio vs. M for gauss markov model for approximation 3

$$\hat{H}_3(\mathbf{R}) = \frac{M}{2}\log(2\pi e) + \frac{M-1}{2}\log(1-\alpha^2) \tag{3.17}$$

Figure 3.10 shows the plot of E_r vs. M. $\hat{H}_3(\mathbf{R})$ is seen to underestimate the joint entropy. Therefore, $\hat{H}_3(\mathbf{R})$ is also a poor entropy estimator for the gauss markov model.

3.4 Summary and Conclusion

In this chapter, three different approximations were analyzed and compared based on the error-ratio (from definition 18). The approximations were tested on the conditionally independent gaussian model and the gauss markov model. The results show that approximation one performs well on both models, approximation two performs well only on the gauss markov model, and approximation three performs poorly on both models. Therefore, it seems that approximation one is a reasonable estimator for the joint entropy.

Chapter 4

Application to Zebra Finch Data

Chechik et. al [4] determined the redundancy in the auditory pathway of cats by applying the redundancy measure r in definition 16 to pairwise neurons. Using data recorded from inferior colliculus (IC), auditory thalamus (MGB), and primary auditory cortex (A1), they found that information about the stimulus identity was slightly reduced in A1 and in MGB neurons in comparison to IC neurons. Therefore, there is a reduction of redundancy in the ascending auditory pathway of cats since neural responses in higher regions have more correlation. Similar to their analysis, in this chapter, we apply the redundancy measure r from definition 16 to measurement data from the auditory system of Zebra Finch Song Birds. However, instead of applying the redundancy measure to pairwise neurons, we apply it to populations of neurons. Data recorded from the midbrain nucleus, mesencephalicus lateralis dorsalis (Mld) and the primary forebrain region, Field L are used in the analysis. In the Zebra Finch auditory system, Mld receives information from multiple lower brainstem auditory nuclei and provides information to the auditory thalamus. Field L, the avian thalamo-recipient, transmits information to the song nuclei, which responds with high selectivity to the sound of the birds own song. The goal of this chapter is to characterize and to distinguish between these two different regions based on their relative amounts of redundancies.

The stimuli used for the experiments are conspecific songs. Experiments show that these

songs elicit responses from the Zebra Finch Birds. Figure 4.1 shows the spectrogram (time frequency plot) of a conspecific stimulus. The presence of the harmonic stacks at specific times shows that there exists time and frequency correlation in the stimulus.

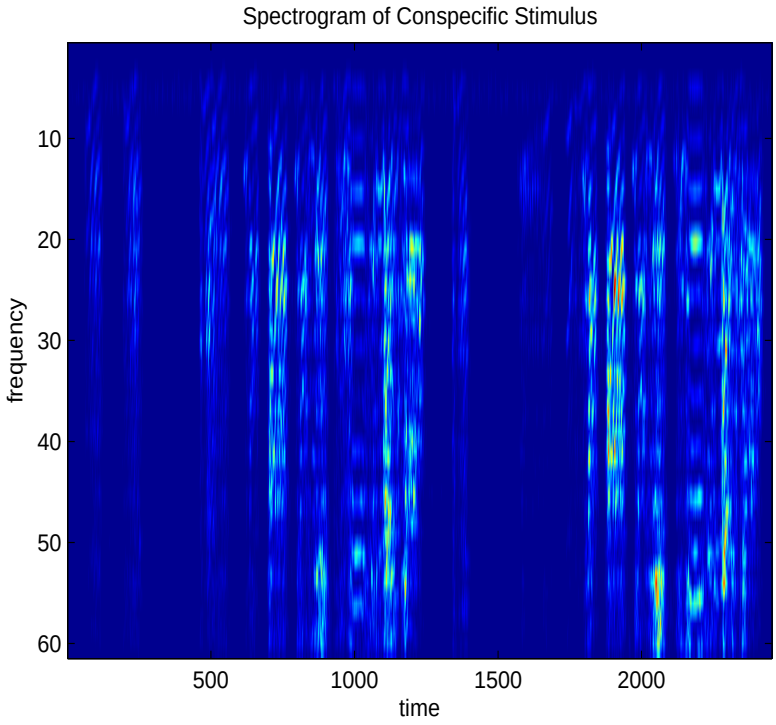


Figure 4.1. spectrogram of conspecific stimulus

The neural responses come from single unit recordings. Figure 4.2 gives a single trial of the response of a neuron to the conspecific stimulus in Figure 4.1. The low number of spikes confirms that there is limited neural data.

Figure 4.3 shows plots of redundancy r vs. number of neurons M for Mld and field L. For each given M, the redundancy r is obtained from averaging fifteen different repeats. A window size of 30 ms was used (see appendix).

Another information measurement of interest is the average information each neuron in the population transmits about the stimulus. This information measurement provides a means to verify the redundancy of a population.

Definition 19. *Given stimulus S, and population response $R_1, R_2, ..., R_M$, the transmission*

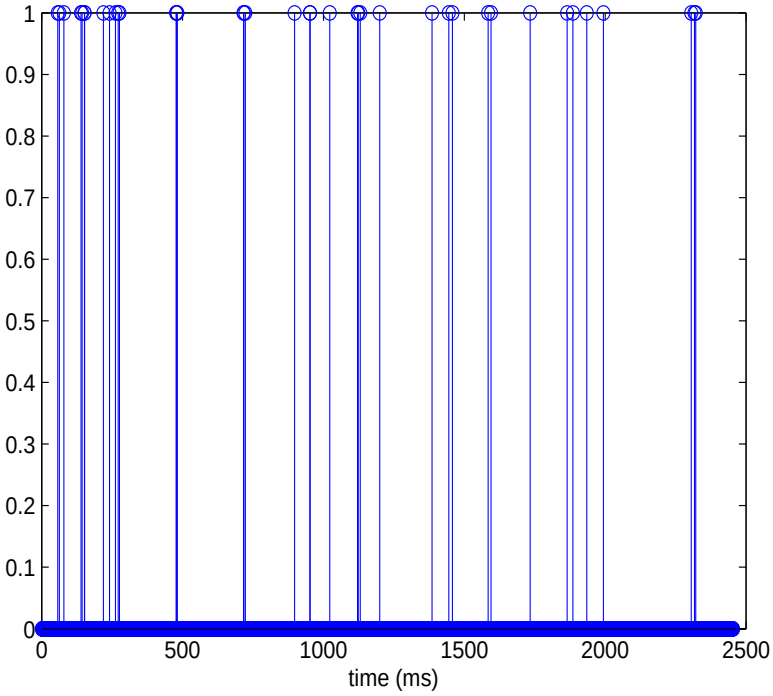


Figure 4.2. spiking response of actual neuron

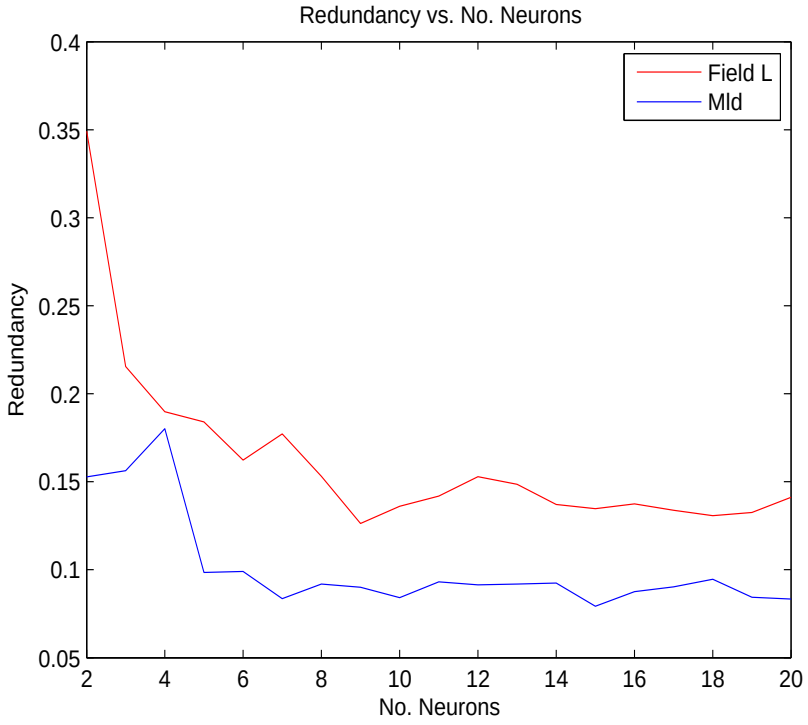


Figure 4.3. redundancy (r) vs. no. neurons (M) for Mld and field L

value (T) of the given region is defined as

$$T = \frac{\sum_{i=1}^{M} I(S; R_i) - I(R_1; R_2; ...; R_M)}{M} \tag{4.1}$$

In a redundant population, the joint information $I(R_1; R_2; ...; R_M)$ is high due to strong correlation between the responses $R_1, R_2, ..., R_M$. From definition 19, high joint information between the population response causes the average information transmitted about the stimulus by each neuron in the population to be low. This aligns with the intuition that in a redundant population, neurons often contain the same information about the stimulus. Figure 4.4 contains plots of transmission values vs. number of neurons for Mld and Field L.

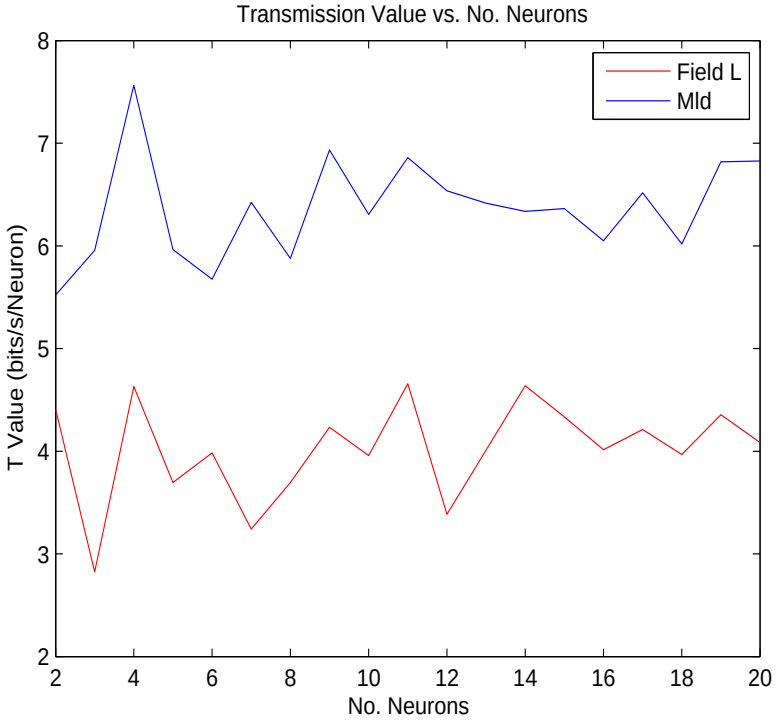


Figure 4.4. Transmission Values (T) vs. no. neurons (M) for Mld and field L

4.1 Discussion

The main goal of this chapter is to use information measures to distinguish between neural regions in the auditory system of song birds. Using data from Mld and Field L, the

simulations show that the two different information measures can be used to characterize the neural regions. Figure 4.3 shows that the first information measure, redundancy, is different for the two neural regions. Field L has a redundancy of approximately 15 percent and Mld has a redundancy of approximately 8 percent, half of that of Field L. Similarly, the second information quantity considered in this thesis, the transmission value, also nicely identifies the two brain regions. Figure 4.4 shows that Field L transmits roughly 6.5 bits per neuron while Mld transmits roughly 4 bits per neuron.

Interpreting the results in Figures 4.3 and 4.4 is slightly more difficult. One hypothesis suggests that redundancy should decrease in higher processing regions of the brain, resulting in more efficient representation and coding of information. This hypothesis has been confirmed in a recent literature [4], in which the reduction of redundancy in the ascending auditory pathway of cats was discovered.

However, simulation results in this chapter suggests the converse of the previous hypothesis to be true. Figure 4.3 shows that Field L is more redundant than Mld. Therefore, redundancy increases in the higher neural region of the auditory system. This result may be caused by a number of reasons. First, we could be systematically oversampling from one corner of Field L with similar receptive fields. Therefore, the neurons are all sensitive to the same type of stimulus and there is high correlation between the responses. Second, our noise and signal models may not be correct. Third, Mld and Field L may have different coding strategies. The following hypothesis based on information theory explains this reason in more detail.

From an information theory point of view, redundancy in a given neural region is determined by the region's coding constraints, which facilitate computation. For example, given a neural region with M neurons and no coding constraints, it follows that a neuron's response to the stimulus is independent of the response of other neurons. Hence, 2^M different representations are possible. On the contrary, consider a region of M 'grandmother cells' where only one neuron can spike in response to the stimulus. With this strict coding constraint, only M representations are possible. In addition, the population is redundant since all except one neuron remain silent. From these two examples, it can be seen that

information theory suggests another hypothesis of information processing in neural systems. This hypothesis implies that different regions have different coding constraints, leading to different redundancies. Therefore, higher regions may not code less redundantly as a result of tighter coding constraints.

4.A Methods

To calculate the redundancy r for a population of neurons, it is necessary to find

1. $H(R_1, R_2, ..., R_M)$
2. $H(R_i)$ for all $i \in \{1, 2, ..., M\}$
3. $I(S; R_i)$ for all $i \in \{1, 2, ..., M\}$

Approximation $\hat{H}_1$ is to to calculate (1). Therefore, for all i, j, the following distributions have to be estimated from data,

1. $p(R_i)$
2. $p(R_i, R_j)$ for all $i, j \in \{1, 2, ..., M\}$.
3. $p(S; R_i)$ for all $i \in \{1, 2, ..., M\}$

Notation

- S: stimulus (in spectrogram form)
- $f \times n$: dimensions of stimulus
- w: window size
- S_i: stimulus block from time $w(i-1)$ to wi
- $\mathbf{R_i}$: response i
- T: total number of response trials to stimulus S

Figure 4.5. Concatenated Response $\hat{R}_i$.

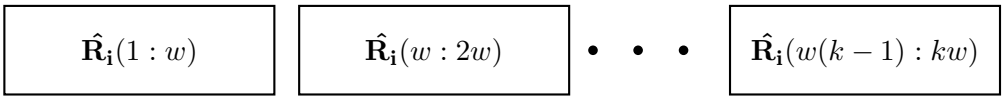


Figure 4.6. Division of Concatenated Response into Blocks.

- $R_i^{(j)}$: trial j of response i
- $\hat{R}_i$: concatenated response i, $(R_i^{(1)}, R_i^{(2)}, ..., R_i^{(k)})$
- k: number of blocks $\frac{n}{w}$

1. Estimating $p(R_i)$

For neuron i, the concatenated response $\hat{R}_i$ is formed by joining the T response trials into one long row vector as shown in figure 4.5. The concatenated response $\hat{R}_i$ is separated into block of length w: $\hat{R}_i(1:w)$, $\hat{R}_i(w:2w)$,...,$\hat{R}_i(w(k-1):wk)$. This process is shown in figure 4.6. The number of spikes in each block is used to form the probability distribution $p(R_i)$.

2. Estimating $p(R_i, R_j)$

For neurons i, j, the concatenated responses $\hat{R}_i$ and $\hat{R}_i$ are separated into blocks of length w: $\hat{R}_i(1:w)$, $\hat{R}_i(w:2w)$,...,$\hat{R}_i(w(k-1):wk)$ and $\hat{R}_j(1:w)$, $\hat{R}_j(w:2w)$,...,$\hat{R}_j(T(w-1):Tw)$. The spike counts in each corresponding pair of blocks $\hat{R}_i(w(c-1):wc)$, $\hat{R}_j(w(c-1):wc)$ are used to determine the probability distribution $p(R_i, R_j)$.

3. Estimating $p(S, R_i)$

The stimulus is separated into blocks of length w: S_1, S_2,...,S_T. Similarly, the response $\mathbf{R_i}$ is separated into blocks of length w: $\mathbf{R_i}(1:w)$, $\mathbf{R_i}(w:2w)$,..., $\mathbf{R_i}(w(k-1):wk)$. The stimulus blocks $S_1, S_2, ..., S_k$ are assumed to be i.i.d. Therefore, for each $c \in \{1, 2, ..., k\}$, the response block $\mathbf{R_i}(w(c-1):wc)$ corresponds to stimulus block S_c. The number of spikes in each trial of the response block $\mathbf{R_i}(w(c-1):wc)$ is used to determine the probability distribution $p(R_c|S_c)$. That is, the number of spikes in each $R_i^{(1)}(w(c-1):wc)$, $R_i^{(2)}(w(c-1):wc)$,...,$R_i^{(T)}(w(c-1):wc)$ is used to determine $p(R_c|S_c)$. Finally, by assuming that the stimulus is uniformly distributed over $\{S_1, S_2, ..., S_k\}$, the distribution $p(S, R_i)$ given as

$$p(S, R_i) = \sum_{c=1}^{k} p(R_c|S_c)\frac{1}{k} \tag{4.2}$$

4. Window Size

In estimating the mutual information between stimulus and response $I(S; R_i)$, it is assumed that the stimulus is divided into i.i.d blocks of window size w. Another method of estimating mutual information is by modeling the spiking neurons as gamma processes [2]. We assumed that the gamma approach gives a good estimate of mutual information. Figure 4.7 compares the mutual information calculated assuming i.i.d stimulus blocks with mutual information calculated assuming gamma model for different window sizes. It can be seen that for window sizes of 30 and 50 ms, the two mutual informations correspond closely. Therefore, a window size of 30 ms was chosen in estimating the quantities $I(S; R_i)$, $H(R_i)$ and $H(R_1, R_2, ..., R_M)$.

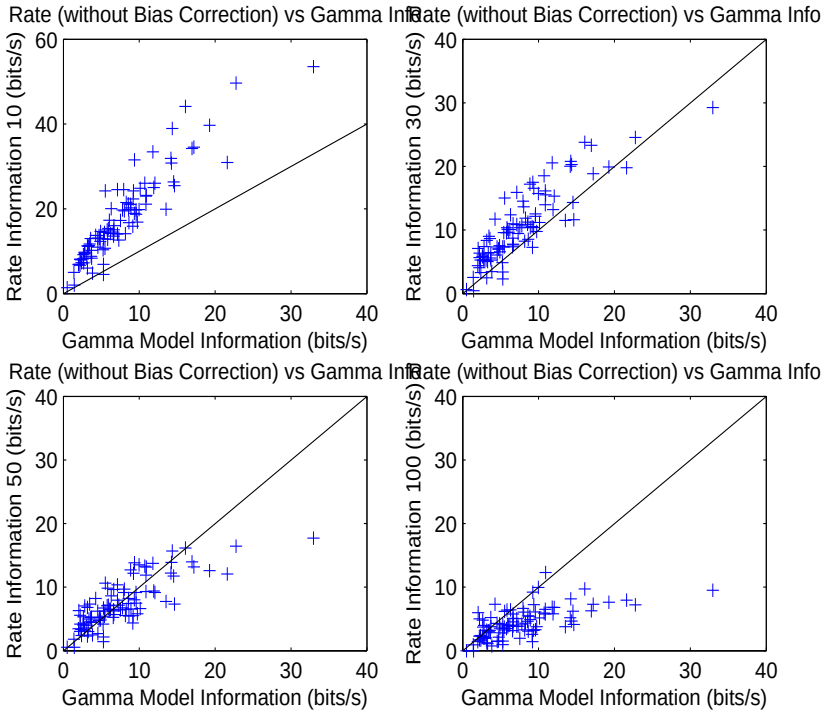


Figure 4.7. Comparison of Mutual Information

Chapter 5

An Alternative Approach to Redundancy

Definition 17 (from chapter 1) implies that population redundancy can be characterized by the mutual information between the stimulus and the population response $I(S; R_1, R_2, ..., R_M)$. However, an accurate estimate of the high dimensional probability distribution $p(S, R_1, R_2, ..., R_M)$ is necessary to obtain a precise calculation of the mutual information. This is difficult in the presence of limited neural data. The goal of this chapter is to present a coarse characterization of population redundancy via describing the scaling behavior of $I(S; R_1, R_2, ..., R_M)$ in M.

5.1 Simple Examples

In this section, the mutual information between stimulus and population response $I(S; R_1, ..., R_M)$ will be characterized as a function of M for a series of simple examples.

5.1.1 Gaussian Stimulus through Gaussian Channel

First, the conditionally independent gaussian model from section 2.1 is analyzed. Since the neural responses come from single unit recordings, they can be assumed to be condi-

tionally independent given the stimulus. The mutual information between stimulus and population response evaluates to

$$I(S; R_1, R_2, ..., R_M) = \frac{1}{2}\log(1 + M\frac{\sigma^2}{\sigma_w^2}) \tag{5.1}$$

Therefore, for large M, it follows that

$$I(S; R_1, R_M, ..., R_M) \sim \log M \tag{5.2}$$

Since $\frac{d\log M}{dM} \to 0$, for M large, an additional neuron provides negligible new information about the stimulus. This is consistent with the intuition that there is an inherent degree of redundancy in the responses $R_1, ..., R_M$ since they are jointly dependent through the stimulus. In addition, the information in the stimulus is limited in the first place since $H(S)$ is bounded. Therefore, when the population is large, additional neurons will not increase the information in the population response.

5.1.2 Binary Stimulus through Binary Symmetric Channel

This second model assumes that the stimulus and the responses are binary. This example captures the spiking nature of the responses.

Definition 20. *Let $S \sim B(\frac{1}{2})$. Let $W_1, W_2, ..., W_M$ be i.i.d $B(\epsilon)$ for $0 < \epsilon < 1$. Each response $R_i = S \oplus W_i$ where $\oplus$ is the modulo two sum. Therefore, R_i is the result of S passed through a binary symmetric channel with cross over probability ϵ as shown in figure 5.1. It follows that $R_1, R_2, ..., R_M$ are bernoulli random variables conditionally independent given S.*

Definition 21. *Let $R_1, R_2, ..., R_M$ be bernoulli random variables. The majority vote $\Lambda(\mathbf{R})$ where $\mathbf{R} = R_1, R_2, ..., R_M$ is defined as*

$$\Lambda(\mathbf{R}) = 1(\sum_{i=1}^{M} R_i \geq \frac{M}{2}) \tag{5.3}$$

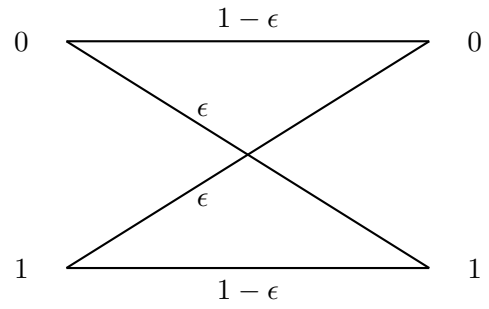


Figure 5.1. Binary Symmetric Channel.

It can be shown that $\Lambda(\mathbf{R})$ is a sufficient statistic of R. Therefore, S and $\mathbf{R}$ are conditionally independent given $\Lambda(\mathbf{R})$.

By the chain rule of mutual information, $I(S; \mathbf{R}, \Lambda(\mathbf{R}))$ can be simplied to be

$$I(S; \mathbf{R}, \Lambda(\mathbf{R})) = I(S; \Lambda(\mathbf{R})) + I(S; \mathbf{R}|\Lambda(\mathbf{R})) \tag{5.4}$$

$$= I(S; \Lambda(\mathbf{R})) \tag{5.5}$$

Similarly, from the conditional independence of S and $\mathbf{R}$ given $\Lambda(\mathbf{R})$, it follows that

$$I(S; \mathbf{R}, \Lambda(\mathbf{R})) = I(S; \mathbf{R}) + I(S; \Lambda(\mathbf{R})|\mathbf{R}) \tag{5.6}$$

$$= I(S; \mathbf{R}) \tag{5.7}$$

Therefore,

$$I(S; \mathbf{R}) = I(S; \Lambda(\mathbf{R})) \tag{5.8}$$

Definition 22. *For this bernoulli example, the maximum likelihood estimate $\hat{S}$ of stimulus S given responses $R_1, R_2, ..., R_M$ is*

$$\hat{S} = \arg \max_{s \in \{0,1\}} p(R_1, R_2, ..., R_M|s) \tag{5.9}$$

and the probability of error P_e associated with this estimate is defined to be

$$P_e = p(\hat{S} \neq S) \tag{5.10}$$

It can be shown that the maximum likelihood test simplifies to

$$\hat{S} = \Lambda(\mathbf{R}) \tag{5.11}$$

and $\Lambda(\mathbf{R}) \sim B(\frac{1}{2})$. Therefore, it follows that

$$I(S;\mathbf{R}) = H(\Lambda(\mathbf{R})) - H(\Lambda(\mathbf{R})|S) \tag{5.12}$$

$$= 1 - H(P_e) \tag{5.13}$$

$$\leq 1 - P_e \tag{5.14}$$

For large M, $H(P_e) \approx P_e$. Hence, the bound in the last step becomes tight.

The Chernoff Bound [5] states that the error exponent is given by the KL divergence between the stimulus distribution and the noise distribution. That is,

$$\log P_e = D(\frac{1}{2}||\epsilon) \tag{5.15}$$

Therefore, the mutual information $I(S;\mathbf{R})$ can be approximated by

$$I(S;\mathbf{R}) \approx 1 - 2^{MD(\frac{1}{2}||\epsilon)} \tag{5.16}$$

As the number of neurons increase, the mutual information between stimulus and population response approaches one bit exponentially fast. This is intuitive since $H(S) = 1$. Thus, the maximum information in the stimulus is one bit. This characterization of information states that for large populations, an additional neuron will give negligible new information about the stimulus.

5.1.3 Bernoulli Stimulus with Choosing Probability

In both of the previous examples, the responses $R_1, R_2, ..., R_M$ are identically distributed. That is, $p(R_1) = p(R_2) = ... = p(R_M)$. Furthermore, for every neuron, the probability distribution of its response conditioned on the stimulus $P(R_i|S)$ is the same as other neurons in the population. In actual neural systems, different neurons may be

sensitive to different realizations of the stimulus . Therefore, these models may have limited impact in characterizing encoding of neural systems. In this section, a model in which different neurons react to different realizations of the stimulus is presented.

Notation:

- $\mathbf{S}$: vector stimulus (S_1, S_2)
- $\mathcal{S}$: set of all realizations of S
- $\mathcal{P}(\mathcal{S})$: power set of $\mathbf{S}$
- P_c: choosing probability
- F_i: receptive field for neuron i
- R_i: response of neuron i
- $\mathbf{R}$: population response $(R_1, R_2, ..., R_M)$

Definition 23. *A neuron's receptive field is the feature in the stimulus which elicits a response in the neuron.*

Definition 24. *The stimulus* $\mathbf{S} = (S_1, S_2)$ *where* S_1, S_2 *i.i.d* $\sim \mathcal{B}(\frac{1}{2})$. *The set of all realizations of the stimulus* $\mathcal{S} = \{(0,0), (0,1), (1,0), (1,1)\}$. *The 'choosing probability'* P_c *is a probability distribution over the set of all realizations* $\mathcal{S}$. *The receptive fields for the neurons* $F_1, F_2, ..., F_M$ *are modeled as i.i.d random variables each with probability distribution* P_c. *Therefore, for all* i, *for all* $\mathbf{s} \in \mathcal{S}$, $p(F_i = \mathbf{s}) = P_c(\mathbf{s})$. *Each response,* $R_i = 1\{F_i = \mathbf{S}\}$. *Therefore, each response chooses a receptive field independently from other responses and reacts if the stimulus corresponds with its chosen receptive field.*

Unlike the previous models, a direct characterization of the mutual information between stimulus and population response $I(S; \mathbf{R})$ is difficult. Therefore, the expectation over the choosing probability of the mutual information $E_{P_c}[I(\mathbf{S}; \mathbf{R})]$ is analyzed in order to give a coarse characterization of the population redundancy.

Definition 25. *For* $\mathcal{T} \in \mathcal{P}(\mathcal{S})$, $A_{\mathcal{T}}$ *is the event that* $F_1, F_2, ..., F_M$ *take values only in* $\mathcal{T}$.

Example 3. *Let* $\mathcal{T} = \{\{0,0\},\{0,1\}\}$. $A_{\mathcal{T}}$ *is the event that for all* $i \in \{1,...,M\}$, F_i *takes the value* $\{0,0\}$ *or* $\{0,1\}$.

The expected mutual information can then be evaluated to be

$$E_{P_c}[I(\mathbf{S};\mathbf{R})] = E_{P_c}[H(\mathbf{R}) - H(\mathbf{R}|\mathbf{S})] \tag{5.17}$$

$$= E_{P_c}[H(\mathbf{R})] \tag{5.18}$$

$$\tag{5.19}$$

For $\mathcal{T} \in \mathcal{P}(\mathcal{S})$, let $H_{\mathcal{T}}(\mathbf{R})$ be $H(\mathbf{R})$ in the event $A_{\mathcal{T}}$. The expected mutual information then becomes

$$E_{P_c}[I(\mathbf{S};\mathbf{R})] = E_{P_c}[H(\mathbf{R})] \tag{5.20}$$

$$= \sum_{\mathcal{T}\in\mathcal{P}} (\mathcal{S}) P(A_{\mathbf{T}}) H_{\mathbf{T}}(\mathbf{R}) \tag{5.21}$$

For each $\mathcal{T} \in \mathcal{P}(\mathcal{S})$, $H_{\mathcal{T}}(\mathbf{R})$ is evaluated to be

$$H_{\mathcal{T}}(\mathbf{R}) = \begin{cases} H(\frac{1}{4}), & |\mathcal{T}| = 1 \\ H(\frac{1}{4}) + \frac{1}{3}H(\frac{1}{3}), & |\mathcal{T}| = 2 \\ H(\frac{1}{4}) + \frac{1}{3}H(\frac{1}{3}) + \frac{1}{2}H(\frac{1}{2}), & |\mathcal{T}| = 3,\ |\mathcal{T}| = 4 \end{cases} \tag{5.22}$$

For $\mathcal{T} \in \mathcal{P}(\mathcal{S})$, for $|\mathcal{T}| = i$, let $\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_i}$ be the elements of $\mathcal{T}$. For $|\mathcal{T}| = 1$, the probability of $A_{\mathcal{T}}$ is evaluated to be

$$P(A_{\mathcal{T}}) = P_c^M(\mathbf{s_1}) \tag{5.23}$$

For $|\mathcal{T}| = 2$,

$$P(A_{\mathcal{T}}) = \sum_{i=1}^{M-2} \sum_{j=1}^{M-i-1} \left(\frac{M}{i}\right) \left(\frac{M-i}{j}\right) P_c^i(\mathbf{s_1}) P_c^j(\mathbf{s_2}) P_c^{M-i-j}(\mathbf{s_3}) \tag{5.24}$$

For $|\mathcal{T}| = 3$,

$$P(A_{\mathcal{T}}) = \sum_{i=1}^{M-2} \sum_{j=1}^{M-i-1} \left(\frac{M}{i}\right) \left(\frac{M-i}{j}\right) P_c^i(\mathbf{s_1}) P_c^j(\mathbf{s_2}) P_c^{M-i-j}(\mathbf{s_3}) \tag{5.25}$$

For $|\mathcal{T}| = 4$,

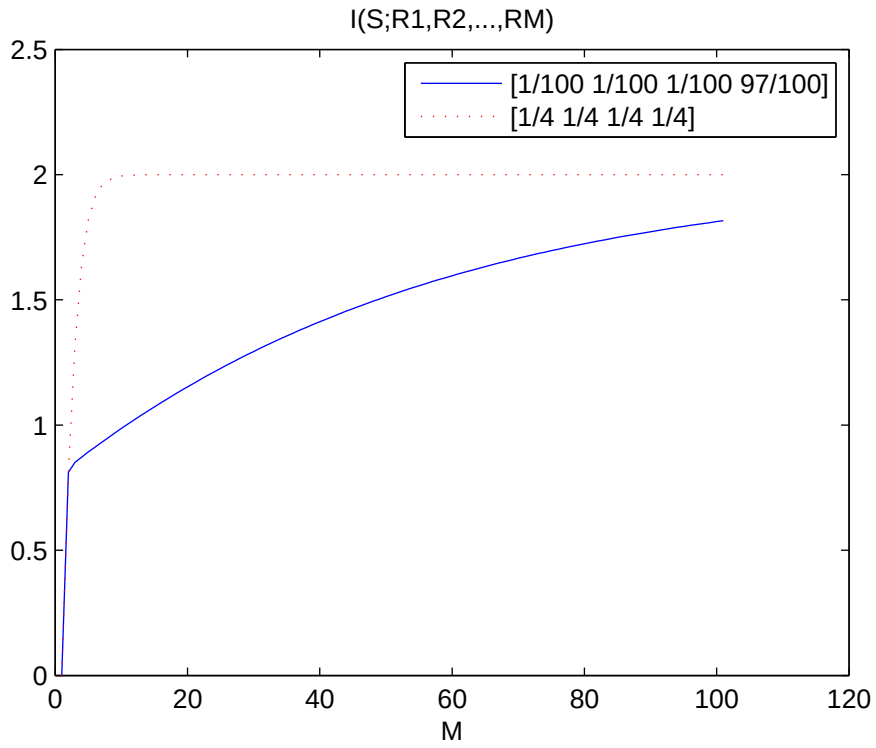
$$P(A_{\mathcal{T}}) = \sum_{i=1}^{M-2} \sum_{j=1}^{M-i-1} \sum_{k=1}^{M-i-j-1} \left(\frac{M}{i}\right) \left(\frac{M-i}{j}\right) \left(\frac{M-i-j}{k}\right) \times \tag{5.26}$$
$$P_c^i(\mathbf{s_1}) P_c^j(\mathbf{s_2}) P_c^{M-i-j}(\mathbf{s_3}) P_c^{M-i-j-k}(\mathbf{s_4})$$

Figure 5.1.3 plots $E_{P_c}[I(\mathbf{S};\mathbf{R})]$ vs. M for two different neural populations. In population one, the responses have choosing probabilities denoted by P_1 where all realizations of the stimulus are equally likely. That is, $P_1(0,0) = P_1(0,1) = P_1(1,0) = P_1(1,1) = \frac{1}{4}$. For population two, the responses have choosing probability P_2 where one of the realization have very high probability while the other three have very low probability. The choosing probability is such that $P_2(1,1) = \frac{97}{100}$ and $P_2(0,0) = P_2(0,1) = P_2(1,0) = \frac{1}{100}$. It can be seen that for both populations, the information approaches the information in the stimulus (two bits) exponentially fast. However, population one approaches at a much faster rate than population two. This suggest that given M neurons, population one will give more information than population two. Therefore, population two is more redundant. The choosing probability P_2 causes most of the neurons in population two to have the same receptive field realization. Therefore, with high probability, additional neurons will not give new information about the stimulus.

5.2 Information Upper Bound

The goal of this section is to present an upper bound of mutual information for a generalized model. The main assumption is that a neuron's response at time t is only dependent on the stimulus from time $t - w$ to t for some window size w. That is, given stimulus $S(t)$, neural response $R(t)$ is only dependent on $S(\tau)$ for $\tau \in [t - w, t]$.

Notation



- S: stimulus (in spectrogram form)
- $\mathcal{S}$: set of all realizations of stimulus
- $f \times n$: dimensions of S
- M: number of neurons
- F_i: receptive field of neuron i
- w: width of receptive field
- X_i: $f \times w$ block of stimulus from time $w(i-1)$ to wi
- R_i: response of neuron i (from time 1 to time n)
- $R_i(a:b)$: R_i from time a to time b
- $\mathbf{R}$: population response $(R_1, R_2, ..., R_M)$
- $\mathbf{R}(a:b)$: population response from time a to time b $(R_1(a:b), R_2(a:b), ..., R_M(a:b))$
- $\mathcal{X}$: set consisting of all receptive field realizations, each of size $f \times w$.

- P_c: choosing probability (distributed over $\mathcal{X}$)
- A: event that S can be reconstructed perfectly from responses $\mathbf{R}$

Definition. *$X_1, ..., X_{\frac{n}{w}}$ are i.i.d stimulus blocks. Note that $S = (X_1, X_2, ..., X_{\frac{n}{w}})$. The receptive fields $F_1, F_2, ..., F_M$ are i.i.d random variables with distribution P_c. Therefore, for all i, for all $x \in \mathcal{X}$, $p(F_i = x) = P_c(x)$. Let the function $g : \mathcal{S} \times \mathcal{X} \to \{0,1\}^n$. Each response, $R_i = g(S, F_i)$. Therefore, each response chooses a receptive field independently from other responses and reacts according to the stimulus and its chosen receptive field.*

The mutual Information between the stimulus and population response is evaluated to be

$$I(S;\mathbf{R}) = I(X_1, X_2, ...X_{n/w}; \mathbf{R}) \tag{5.27}$$

$$= \sum_{i=1}^{n/w} I(X_i; \mathbf{R}|X^{i-1}) \tag{5.28}$$

$$= \sum_{i=1}^{n/w} I(X_i; \mathbf{R}, X^{i-1}) - I(X_i; X_{i-1}) \tag{5.29}$$

Since X_i are i.i.d, $I(X_i; X_{i-1}) = 0$. Therefore,

$$I(S;\mathbf{R}) = \sum_{i=1}^{n/w} I(X_i; \mathbf{R}, X^{i-1}) \tag{5.30}$$

$$= \sum_{i=1}^{n/w} I(X_i; \mathbf{R}) + I(X_i; X^{i-1}|\mathbf{R}) \tag{5.31}$$

$$\tag{5.32}$$

However, since the receptive fields are of width w, only the block $\mathbf{R}(w(i-1) : wi+w)$ of the population stimulus gives information on X_i. From X^{i-1}, only X_{i-1} affects $\mathbf{R}(w(i-1) : wi + w)$. Therefore, X_i and X^{i-2} are conditionally independent given $\mathbf{R}$ and X_{i-1}. The mutual information then simplifies

$$I(S;\mathbf{R}) = \sum_{i=1}^{n/w} I(X_i; \mathbf{R}) + I(X_i; X_{i-1}|\mathbf{R}) \tag{5.33}$$

Recall that distribution P_c is the choosing probability, and A is the event that the stimulus S can be reconstructed perfectly from population response $\mathbf{R}$. Let $I_A(S;\mathbf{R})$ denote $I(S;\mathbf{R})$ when A occurs. The normalized expected mutual information evaluates to be

$$\frac{1}{n}E_{P_c}[I(S;\mathbf{R})] = \frac{1}{n}P(A)I_A(S;\mathbf{R}) + \frac{1}{n}P(A^c)I_{A^c}(S;\mathbf{R}) \tag{5.34}$$

$$\leq \frac{P(A)}{n}I_A(S;\mathbf{R}) \tag{5.35}$$

Using 5.33, the expected mutual information becomes

$$\frac{1}{n}E_{P_c}[I(S;\mathbf{R})] \leq \frac{P(A)}{n}\left(\sum_{i=1}^{n/w} I_A(X_i;\mathbf{R}) + I_A(X_i;X_{i-1}|\mathbf{R})\right)$$

Since A is the event that the stimulus S can be reconstructed perfectly from population response $\mathbf{R}$, for all i, $H_A(X_i|\mathbf{R}) = 0$. Therefore, $I_A(X_i;X_{i-1}|\mathbf{R}) = 0$, and

$$\frac{1}{n}E_{P_c}[I(S;\mathbf{R})] \leq \frac{P(A)}{n}\sum_{i=1}^{n/w} I_A(X_i;\mathbf{R}) \tag{5.36}$$

$$\leq \frac{P(A)}{n}\sum_{i=1}^{n/w} H(X_i) \tag{5.37}$$

Since X_i are i.i.d

$$\frac{1}{n}E_{P_c}[I(S;\mathbf{R})] \leq \frac{P(A)}{n}\frac{n}{w}H(X) \tag{5.38}$$

$$\leq \frac{P(A)}{w}H(X)$$

Therefore, the expected mutual information between stimulus and response is upper bounded by $\frac{P(A)}{w}H(X)$. For fixed w and X, this bound depends on $P(A)$, the probability that all receptive fields are chosen by the neural population. Since $P(A)$ is governed by the choosing probability P_c, the mutual information can be characterized by P_c. For large M,

this bound is tight since for all P_c, $P(A^c) \to 0$ as $M \to \infty$. Therefore, this model suggests that if different neural regions have distinct choosing probabilities, then they can can be distinguished by their respective redundancies.

References

[1] Glovazky A. Determination. of redundancies. in a set of patterns. *Information Theory, IEEE Transactions*, IT-2:151–153, 1956.

[2] Hsu A.S. *Neural encoding of natural sounds in the auditory system.* PhD thesis.

[3] Zee A. Bialek W. Coding and computation with neural spike trains. *Journal of Statistical Physics*, 59:103–115, 1990.

[4] Bar-Yosef O. Young E. Tishby N. Nelken I. Chechik G., Anderson M. Reduction of information redundancy in the ascending auditory pathway. *Neuron*, 51(3):359–368, 2006.

[5] Thomas J.A Cover T.M. *Elements of Information Theory.* Wiley-Interscience, 1991.

[6] Laubach M. Narayanan N.S, Kimchi E.Y. Redundancy and synergy of neuronal ensembles in motor cortex. *J Neuroscience*, 25:4207–4216, 2005.

[7] Harris R.A Berry M.J II Puchalla J., Schneidman E. Redundancy in the population code of the retina. *Neuron*, 46:493–504, 2005.

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

[9] Berry M.J II Schneidman E., Bialek W. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23:12539–12553, 2003.

[10] Bialek W. Doupe A.J Wright B.D, Sen K. Spike timing and the coding of naturalistic sounds in a central auditory area of songbirds. *Advances in Neural Information Processing*, 14:309–316, 2002.