# An Information Theoretic Analysis of Neural Multiplexing

Ezekiel Williams

Thesis submitted to the Faculty of Science in partial fulfillment of the requirements
for the degree of
Master of Science Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa



[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

# Abstract

How the brain encodes information in sequences of voltage spikes is an open question. Past literature suggests the importance of bursts, high-frequency spike events, as a key step towards answering this question. In particular, it was recently shown that neurons could use bursts to communicate two streams of information simultaneously, resulting in higher information rates than seen with other neural code theories. However, it is unknown how a neuron's spiking statistics might affect communication via this new code. To investigate the influence of spike statistics, we study a bursting neuron model with the goal of estimating its information rate as a function of its spike statistics. To this end we extend a recently proposed method for estimating information rate. We find the information rate in our burst-multiplexing model is robust to changes in spike-train statistics, providing evidence for the utility of a burst-multiplexing code to diverse brain networks.

# Résumé

Comment le cerveau encode l'information sous la forme de séquences de décharges électriques est une question ouverte. La littérature scientifique suggère l'importance des bouffées ("bursts"), des décharges à haute fréquence, comme étant une clé de la réponse à cette question. En particulier, il a été démontré récemment que des neurones pouvaient utiliser les bouffées pour communiquer simultanément deux sources d'information, entraînant des taux d'information transmis plus élevés que ceux observés avec d'autres théories du code neuronal. Cependant, on ne sait pas comment les statistiques de décharge d'un neurone pourraient affecter la communication via ce nouveau code. Pour étudier l'influence de ces statistiques, nous étudions un modèle de neurones produisant des bouffées dans le but d'estimer son taux d'information transmis en fonction de ses statistiques de décharge. À cette fin, nous étendons une méthode récemment proposée pour estimer le taux d'information. Nous trouvons que le taux d'information transmis dans notre modèle de multiplexage via les bouffées est robuste aux changements dans les statistiques de décharge, fournissant une preuve de l'utilité d'un tel code pour divers réseaux du cerveau.

# Dedications

This thesis is dedicated to those who are worried they don't have what it takes but keep hitting their head against the wall until they breakthrough.

# Acknowledgement

I would first like to thank my supervisors, Richard and Maia, without whose kind help and guidance none of this would have been possible. I would also like to thank my friend and fellow lab member Alexandre Payeur for corrections, advice, proof reading, positive outlook and crow research–at least a subset of which was invaluable for the development of this thesis. I would like to thank all the members of the Neural Coding lab for the many discussions and for putting up with me as a lab/office-mate. Outside of the academic world, I thank my roommates at Cozy Concord for keeping me grounded and sane (but crazy) and, as always, my wonderful family who have, to my continued surprise, not disowned me just yet.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 The Neural Code

### 1.1.1 Why Study the Neural Code

Understanding how information is communicated in the nervous system is of immense importance to neuroscience, Artificial Intelligence (AI) and society. The ability to read the neural code would grant physicians insight into the mechanisms of neurological disease, paving the way to novel mental health treatments. A neural codebook would allow engineers to design better Brain Computer Interfaces (BCIs) to explore the secrets of cognition, unlock the advances in communication being sought by companies like Neuralink and Facebook, and create innovative new means of prosthetic control for those with missing limbs or impaired sensory systems. Modern AI has high energy demands [3]—a major problem given the environmental crisis the world is currently experiencing. A complete understanding of the neural code would open the door to effective implementation of machine learning techniques on neuromorphic hardware, which has been shown to reduce energy costs substantially [22]. With so many benefits to reading the neural code it is unsurprising that the study of neural communication has a long history.

### 1.1.2 Birth of Neural Code Theory

Ramón y Cajal's "Neuron Doctrine" [27] theory, proposed near the end of the 19th century, set the stage for our modern understanding of the neuron as the fundamental computational unit of the nervous system. Single neurons are connected to each other by junctions called synapses, where information transmission is, for the extent of the material covered in this thesis, directional. Neurons communicate using "spikes", electrical impulses, and it is standard to refer to the spike-generating neuron as pre-synaptic and the receiving cell as post-synaptic. A single neuron will integrate incoming signals from other neurons until the voltage of its cell membrane reaches a threshold, then emit its own spike to be picked up and integrated by other neurons. These principles have been understood since the early 20th century [51] and, since then, the field of neuroscience has built off of two empirically supported axioms: first, that the transmission and integration of spike patterns across networks of neurons underlies cognition and, second, that only the presence or absence of a spike carries information. Taken together, these imply that the neural code can be understood as being composed of binary time series, or spike trains, where 1 represents the presence of a spike and 0 the absence.

### 1.1.3 Modern Theories of Neural Coding

One can classify the many neural coding theories suggested over the past decades into two categories [9]: rate codes and spike-timing dependent codes. Rate

coding is the idea that the mean rate of spikes, the mean being taken over time windows or over a population of neurons, contains all the information encoded in a spike train. An example of this would be a light sensitive neuron in the retina firing a higher rate of spikes for brighter lights. On the other hand, a spike-timing dependent, or temporal, code is one in which the relative timing of spikes conveys information. Consider a sequence of three spikes in a fixed period of time, occurring in one's brain in response to viewing an image. A purely pedagogical example of a temporal code would be if the occurrence of the middle spike closer to the first spike implied the presence of a cat in the image while the middle spike occurring closer to the third spike implied the presence of a dog. One might relate temporal and rate codes by considering a temporal code to be a rate code that is averaging over an infinitesimally short time window and utilizing spike train features at the level of a single neuron.

Strong support has been shown for both rate and temporal theories. In the pioneering rate code experiment of Adrian & Zotterman [1] the muscle fiber of a frog was stretched to different tensions while recording from the neurons innervating the muscle. It was observed that higher tension resulted in a faster rate of neural spikes. More recently, modern BCIs are able to extract significant amounts of information from neural firing using a rate coding hypothesis alone [55]. In support of a temporal code it has been shown that certain organisms exhibit behaviours that cannot be performed without taking into account precise spike timing [59, 13]. Furthermore, some cells exhibit such sparse spike patterns that estimation of spike rate, a necessity given a rate code, would be impossible without averaging over unrealistically long time windows [79]. This is a powerful argument given the rapid behavioural responses required by humans and other animals [40, 73]. With evidence for both rate and timing-dependent theories, the consensus seems to be that both codes are used by the brain [62].

### 1.1.4 Bursts, ISI Distribution and the Neural Code

A burst is a stereotyped pattern of two or more spikes fired at a much higher frequency than the average spike rate of the given neuron. Neuroscientists have been interested in bursts for several decades because many cells in diverse areas of the brain exhibit bursting [41] and because a body of work suggests the importance of bursts in the neural code. Bursts result in more robust information transmission between neurons than singlet spikes, decreasing the noisiness of neural signals [50]. Bursts have been associated with higher cognitive processes like attention or goal-directed behaviour [76]. Bursts correlate with particular sensory events [14]. For example, bursts occur in neurons of the cat visual pathway when a new object appears in the animal's field of vision [47]. These latter two observations provide evidence for a particular form of temporal code where a burst signifies something different than a single spike.

If spikes make up the alphabet of the neural code then the phrases and sentences of the code are composed of spike trains. One should thus study the statistics of spike trains to better understand neural communication–a statement that is underlined by the significant amount of literature on spike train statistics [60].

An important tool for studying burst firing, and spike train statistics more generally, is the Inter-Spike Interval (ISI). The ISI is the real-valued random variable giving the period of time between successive spikes in a spike train. The ISI distribution is useful for quantifying the extent that a neuron produces bursts as frequently bursting neurons will have a peak in this distribution at small ISI values [6]. Furthermore, the ISI distribution can provide evidence for or against a temporal code by its closeness to the interval distribution of the Poisson process (see section 2.1.2). Intervals of a Poisson process are independent so spike trains generated by such a process contain no information in the relative temporal pattern of their spikes [62]. The high variability in ISI distribution across cells [53, 43, 15, 69, 67] in the nervous system is notable, with distributions typically spanning a spectrum from unimodal to bimodal.

## 1.2 Burst Multiplexing: A New Code Theory

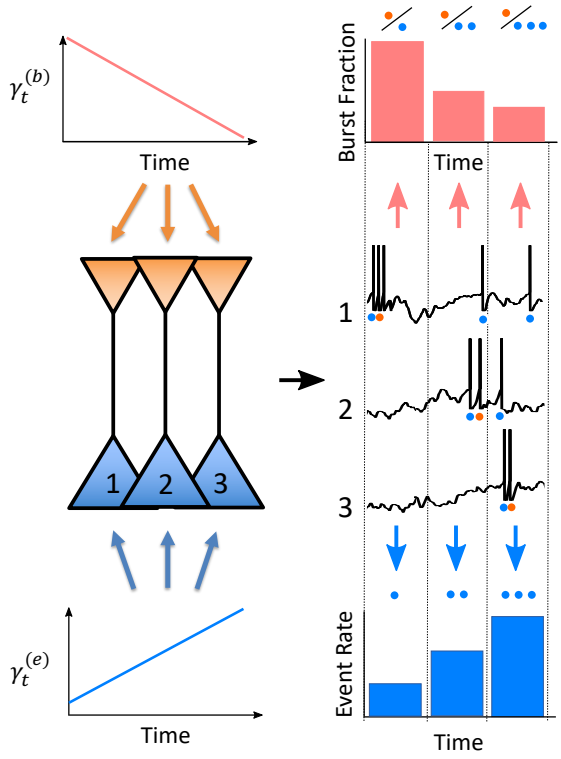


Figure 1.1: **Introduction to Burst Multiplexing:** Signals $\gamma_t^{(b)}$ and $\gamma_t^{(e)}$ are encoded in the spiking of a population composed of cells 1, 2 and 3. The spike-trains produced by these three cells encode the top signal in the burst fraction and the bottom signal in the event rate.

### 1.2.1 Intro to Burst Multiplexing

Naud & Sprekeler's recently proposed burst multiplexing theory of neural coding [54] builds upon the idea that bursts might be coding a different kind of information than standard spikes in a spike train (see section 1.1.4). The novelty of this theory is that it provides a framework for understanding how information can be both encoded and decoded using a language of bursts and regular spikes and, in doing so, how the spike trains of a single population of neurons could be used to communicate two signals at once.

In discussing burst multiplexing it is necessary to make distinctions between spikes based on their membership in a burst. To this end, we will adopt the following naming scheme throughout this thesis:

**Definition 1.2.1.** ***Burst spike:*** *a spike within a burst other than the first spike of the burst.*

**Definition 1.2.2.** ***Event spike:*** *a spike that occurs either outside a burst or as the first spike of a burst.*

**Definition 1.2.3.** ***Isolated spike:*** *a spike that occurs outside a burst.*

**Definition 1.2.4.** ***Event:*** *either an isolated spike or a burst. An entire burst is viewed as a single event, regardless of the number of spikes that compose it.*

The definitions for burst and event are those used by Naud & Sprekeler. In this thesis a further distinction is made between 'burst' and 'event' spike because of their distinct roles in the decoding process. These novel definitions reflect the decoding neuron's interpretation of a spike train, as will become clear in the discussion of neural plasticity rules below. Finally, it bears mentioning that the definitions of burst and event used here are chosen primarily to simplify discussion and analysis of the burst multiplexing code. It is likely that what constitutes a burst or an event biologically could be more subtle.

Short-Term Plasticiy (STP), a neural adaptation mechanism, is central to the decoding stage of burst multiplexing. STP is a phenomenon whereby the membrane voltage at a neuron's synapse responds differently to an incoming spike depending on the pattern of previous spikes impinging on the neuron and can thus be thought of as a kind of synaptic memory. STP can be subdivided into Short-Term Depression (STD), and Short-Term Facilitation (STF). STD occurs when a post-synaptic neuron exhibits less of a voltage response to stimulation following a pre-synaptic spike, thus situating a synapse employing STD to miss burst spikes, picking up only event spikes. STF occurs when a post-synaptic cell exhibits a greater response to a spike that follows quickly after another spike. In this way an STF employing synapse is suited to respond strongly to burst spikes, via the memory of the event spike that proceeds them.

The burst multiplexing code provides a recipe for encoding and subsequently decoding two signals concurrently in the spike trains from a population of statistically equivalent neurons. Burst multiplexing posits that independent signals can be encoded in the event rate and burst fraction respectively of a population of neurons. Population event rate is the number of events emitted in the cell population of interest at a given moment in time and population burst fraction is the ratio of the number of bursts to events in the population at the given time point (figure 1.1). Hereafter we will drop the 'population' portion of the name and refer to these quantities simply as event rate and burst fraction, for brevity. Event rate and burst fraction can be decoded in the membrane potentials of downstream, post-synaptic, neurons by applying STP rules to the spike trains of the encoding population and then averaging the resultant time series over the encoding cells. Burst multiplexing

thus makes use of both temporal and rate codes: a temporal code to decipher bursts from events and a population averaging rate code to estimate event rate and burst rate; the former rate representing one decoded signal and the quotient of the two rates representing the other.

### 1.2.2 Support for Burst Multiplexing

Naud & Sprekeler's work provides strong theoretical support for burst multiplexing. In their paper they demonstrated the efficacy of event rate and burst fraction for encoding independent signals and the ability of post-synaptic neurons to subsequently decode these signals using STP. Furthermore, it was also shown in a toy model that a burst multiplexing code could communicate up to twice as much information per unit time than a rate code.

Beyond the theoretical support, there is a body of experimental research that has found neurological mechanisms capable of implementing burst multiplexing. Layer 5 pyramidal cells, in the mammalian neocortex, exhibit a distinct morphology where the upper, dendritic, and lower, somatic, parts of the neuron receive and integrate incoming spike-trains more or less independently [75]. This would allow for the simultaneous processing of two signals. Events in these cells are generated primarily by the somatic membrane voltage while bursting is dictated by dendritic activity [45]. This cellular machinery is very aptly suited for encoding in burst multiplexing spike-trains. Furthermore, layer 5 pyramidal cells are not the only candidates for burst multiplexing, as many cells exhibit similar bursting behaviour, including neurons in other regions of the cortex [77], thalamus [47] and cerebellum [29]. For decoding, STP rules are well document in circuits across the brain.

Burst multiplexing provides an elegant, biologically plausible framework for understanding how rate and temporal codes might be used together and for the functional significance of bursts as a neural phenomenon. However, there are a number of experimental and theoretical questions that must be addressed to accept or refute burst multiplexing. Foremost among these are, from an experimental perspective, determining which regions of the brain and which cell types could support this code and, from a theoretical standpoint, building an understanding of which spike-train statistics are conducive of this novel code framework. These two questions are addressed by this thesis project.

## 1.3 Research Questions

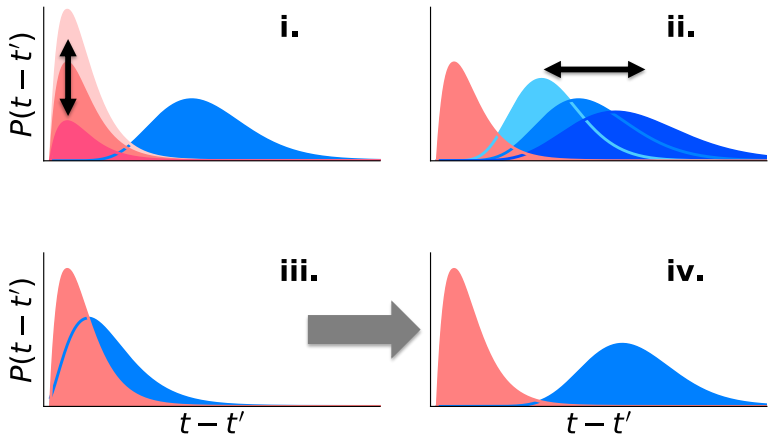


Figure 1.2: **Research Questions: i.** Burst fraction encoded in relative height of intra-burst portion of ISI distribution (orange) **ii.** Event rate encoded in mean of non-intra-burst portion of ISI distribution (blue) **iii-iv.** How does information transmission change as ISI distribution moves from unimodal (iii) to bimodal (iv)?

The high variability in ISI distributions exhibited by neurons (see section 1.1.4) motivates the question of how the burst multiplexing code fares as a function of ISI distribution. Furthermore, ISI distribution is easily measurable and if it can be shown that burst multiplexing is impossible for cells with certain ISI statistics this will allow experimentalists to make statements about which brain regions could exhibit a multiplexed code. An answer to this question of burst multiplexing prevalence could generate insight into the functional significance of bursting in different cell types or, if observed ISI distributions are incompatible with multiplexing, provide evidence against the theory.

Assuming a burst multiplexing code, the ISI distribution is a mixture of intra-burst ISIs, those that appear between spikes in a burst, and all other, longer,

ISIs. Intuitively, one would assume that burst multiplexing should be difficult for unimodal ISI distributions because, since STP rules decode the multiplexed signal using ISI, post-synaptic cells would not be able to distinguish bursts from events. Conversely, forcing the ISI distribution to be bimodal would also influence transmission. To match the power output of a unimodal distribution while keeping two well-separated modes, the variance of at least one mode will have to be decreased. One would assume that this would negatively impact information transmission by reducing the entropy of the output signal or, equivalently, the dynamic range of the encoding cells. Because of these two competing interests we hypothesized that there would be an optimal distribution shape that would maximize information transmission. To test this hypothesis, we first needed a method of modeling burst multiplexing neurons that parameterizes the ISI distribution modality and that captures the biology sufficiently well. Next, we needed a way of quantifying information transmission in networks of neurons employing burst multiplexing.

The above considerations make it clear that a proper analysis of our hypothesis requires an applied math approach. Mathematical modeling is needed in the development of a framework for simulating model neurons; signal processing and probability are necessary for the study of time series and Shannon's information theory is the natural place to find tools for studying communication in a network of neurons.

The current thesis investigates how a burst multiplexing theory of neural coding performs as a function of ISI distribution. The rest of the manuscript will be divided into four parts: background, results, discussion and appendix. The background will introduce and develop the theory employed by the project. The results section will highlight our main contributions to theoretical neuroscience. The discussion will relate these results to the literature and explain future directions in mathematics and in neuroscience. Finally, the appendix will detail the computational methods used along with other supporting material.

# Chapter 2

# Background

To investigate the question of how neural information transmission is affected by ISI distribution shape, given a burst multiplexing code, we first build a model of a simple neural circuit, then analyze the behaviour of the circuit using information theory. The two major subsections of this chapter are Network Model and Estimating Information Rates accordingly. At the start of each of these sections however, we will make short detours to review certain pertinent topics. Before the modeling section we will delve into three topics from probability, on account of the stochastic nature of neural communication. Before the information rates section we will review concepts from signal processing, for their relevance to time series analysis, and, finally, core concepts from information theory.

## 2.1 Three Topics in Probability

The standard definitions [8] of random vector and random variable will be used throughout this thesis.

### 2.1.1 Cumulants

Cumulants are a concept that will be important in the latter parts of the background of this thesis. The cumulants of a probability distribution are a sequence of values that describe the shape of the distribution [8, 11]. Cumulants are closely related to moments in several ways. The first, second and third cumulants correspond to the first, second and third central moments–though cumulants and central moments diverge at and above the fourth order. Regardless of this divergence, if the moments of two distributions are shared then their cumulants are shared also. For this reason cumulants can be used like moments to determine the identity of, or prove useful results about, the associated distribution. Cumulants can be defined using either the natural logarithm of the moment generating function or the characteristic function. Here we will use the former option:

**Definition 2.1.1.** ***Cumulant:*** *Let $X = [X_1, ..., X_n] \in \mathbb{R}^n$ be a random vector. Then the $k^{th}$ order joint cumulant is defined as the coefficient of the $k^{th}$ term of the Taylor expansion of the cumulant generating function:* $\log\left(E(e^{itX})\right)$

We will denote the joint cumulant of a set of $k$, not necessarily unique, random variables $\{X_1, ..., X_k\}$ by $\mathcal{C}[X_1, ..., X_k]$. In this way the $k^{th}$ order cumulant of a univariate random variable $X$ is taken by setting $X_j = X \ \forall j \in \{1, ..., k\}$. Cumulants have several properties that make them easier to deal with than moments. We list the cumulant properties [11] that will be important for this thesis in the following lemma.

**Lemma 2.1.2.** ***Cumulant Properties:*** *Let $X$ be a random variable, $Y = [Y_1, ..., Y_k]$ be a $k$-dimensional random vector and $a = [a_1, ..., a_k] \in \mathbb{R}^n$ be constant. Then*

1. $\mathcal{C}[Y_1, ..., Y_k]$ *is symmetric in the sense that the order of arguments is interchangeable.*
2. *If elements of* $Y$ *can be divided into two independent sets then* $\mathcal{C}[Y_1, ..., Y_k] = 0$
3. $\mathcal{C}[a_1 Y_1, ..., a_k Y_k] = \prod_{j=1}^{k} a_j \mathcal{C}[Y_1, ...., Y_k]$ *(homogeneity)*
4. $\mathcal{C}[a + Y_1, ..., Y_k] = \mathcal{C}[Y_1, ..., Y_k]$ *(translational invariance)*
5. $\mathcal{C}[X + Y_1, ..., Y_k] = \mathcal{C}[X, Y_2, ..., Y_k] + \mathcal{C}[Y_1, ..., Y_k]$ *(additive property)*

### 2.1.2 Stochastic Processes

Stochasticity is vital in this thesis because of its role in neural modeling. The membrane potential of a neuron and the dynamics of synaptic transmission are dictated by an unthinkably large set of variables. To design a deterministic model capturing every small change in the behaviour of a cell would be impossible; instead randomness is used to account for this variability. This rationale combined with the temporal nature of neural processing leads naturally to the study of stochastic processes.

**Definition 2.1.3.** ***Stochastic Process:*** *A stochastic process is defined as a collection of random vectors,* $\{X_t \mid t \in T\}$*, on some shared probability space* $(\Omega, \mathcal{F}, P)$*, where* $T$ *is an index set.*

Stochastic processes can be separated into discrete and continuous varieties, where $T \subseteq \mathbb{Z}$ for the former and $T \subseteq \mathbb{R}$ for the latter. Note that this definition includes stochastic processes that are sequences of univariate random variables if we take the random vector to be one dimensional. We typically describe stochastic processes using their finite dimensional distributions, that is, the joint distribution of a finite selection of values of the stochastic process: $F(X_{t_1}, ...., X_{t_k})$, $k \in \mathbb{N}$.

Notation can get particularly messy when dealing with cumulants of multivariate stochastic processes. We will thus adapt our cumulant notation for dealing with this scenario as follows: let $\left\{X_t \mid t \in T\right\}$ be a stochastic process with some index set $T$ such that $\left\{X_t^{(l)} \mid t \in T\right\}$, $l \in \{1, ..., n\}$ is the process at the $l^{th}$ element of $X$. Let $\mathcal{S}_k = \{l_1, ..., l_k\}$, be some set of not necessarily unique integer indices such that $l_j \in \{1, ..., n\}$, $\forall j \in \{1, ..., k\}$. Thus, $\mathcal{S}_k$ can represent a set of possibly repeating indices of the elements of $X$. We will denote the joint, $k^{th}$ order cumulant of this selection of random variables by $\mathcal{C}_{\mathcal{S}_k}(t_1, ..., t_k) = \mathcal{C}\left[X_{t_1}^{(l_1)}, ..., X_{t_k}^{(l_k)}\right]$.

Three properties of stochastic processes which will play an important part in the information analysis section of the thesis are stationarity, finite memory and ergodicity which are defined below.

**Definition 2.1.4.** ***Stationarity:*** *A stochastic process* $\{X_t \mid X \in \mathbb{R}^n, t \in T\}$ *is strictly stationary if its finite dimensional distributions are shift invariant:* $F(X_{t_1}, ...., X_{t_k}) = F(X_{t_1+\tau}, ...., X_{t_k+\tau})$*, where* $\tau$ *is arbitrary element of* $T$*.*

Two or more stochastic processes are jointly stationary if the multivariate process whose elements are composed of the two or more processes of interest is itself stationary.

**Definition 2.1.5.** ***Finite Memory:*** *A stochastic process* $\{X_t \mid t \in T\}$ *has finite memory if* $\sum_{t_1,...,t_{k-1}=-\infty}^{\infty} \mathcal{C}_{\mathcal{S}_k}(t_1, ..., t_{k-1}, t) = M < \infty \quad \forall t \in \mathbb{Z}$ *and* $k \in \mathbb{N}$ *such that* $k > 1$*.*

Intuitively, finite memory [11] means that the dependence between the random variables at two index points in the stochastic process tends to zero as the separation of the points in the index space gets larger and larger.

**Definition 2.1.6.** ***Ergodicity:*** *A stochastic process is ergodic almost surely if each of its statistics can be calculated from a single, infinitely long, sample path.*

The second order cumulant of a univariate, stationary stochastic process is a commonly used statistic and is referred to as the autocorrelation function, which will be denoted by $R_{XX}$ in this thesis. If $X$ is a stationary stochastic process its autocorrelation function will, by the definition of stationarity, depend only on the difference of the index between the two elements of the stochastic process: $\mathcal{C}(X_{t_1}, X_{t_2}) = \mathcal{C}(X_{t_1-t_2}, X_0) = R_{XX}(\tau)$, $\tau = t_1 - t_2$. The autocorrelation can be generalized to represent the second order cumulant of two stochastic processes, $X$ and $Y$. This statistic, known as the cross correlation, will be denoted $R_{XY}$. If the two processes are jointly stationary then, as in the case of the autocorrelation, the cross-correlation will depend only on the difference in the index of the process elements. Autocorrelation and crosscorrelation will be used in section 2.3.2 to define important frequency domain statistics for spectral processes.

Two classes of stochastic processes, the point process and the Ornstein-Uhlenbeck process, are used to model neural spike-trains and noise, respectively, in this thesis and are reviewed below.

### The Point Process

A point process is a mathematical object describing the random allocation of points to an ambient space. Because we are specifically interested in modeling events occurring in time we have adopted a definition that lends itself to temporal modeling and restricts the ambient space to the real line. However, it bares mentioning that the notion of point process is much more general than this: it can be extended to a variety of ambient spaces and can be viewed from the perspective of a stochastic process and a random measure [64].

**Definition 2.1.7.** ***Point Process:*** *The stochastic process* $\{\mathcal{T}_n \mid n \in N\}$*, where* $N \subset \mathbb{N}$*, is a point process if:*

1. $\mathcal{T}_n \in \mathbb{R} \cup \pm\infty$
2. $\#\{n \mid \mathcal{T}_n \in B\} < \infty$*, where* $B$ *is any bounded Borel set on* $\mathbb{R}$

*We further define the point process to be* ***simple*** *if* $\mathcal{T}_i \neq \mathcal{T}_j$ *a.s.* $\forall i \neq j$

Point processes are important for modeling temporally distributed events, for example the car crashes occurring on a segment of road could be modeled using a point process by letting the real axis represent time and taking $S_n$ to be the time of the $n^{th}$ car crash. In neuroscience point processes are used to model the spike times of neurons.

There are two other perspectives on the point process that are quite important for neural modeling. First, one can consider the sequence of intervals $\{S_{n+1} - S_n \mid n \in N\}$. This can be used to model ISIs in the spike-train of a neuron. Second, one can divide up $\mathbb{R}$ into a sequence of bins $\{[t_i, t_{i+1}) \mid t_i \leq t_{i+1} \forall i \in \mathbb{Z}\}$ and consider the discrete, univariate stochastic process with elements $X_i = \#\{n \mid S_n \in [t_i, t_{i+1})\}$. One frequently bins spike-trains when dealing with neural data and this provides a method of formalizing such a procedure.

We will elaborate on three types of point processes before discussing the Ornstein-Uhlenbeck process.

### Marked Point Process on $\mathbb{R}$

The marked point process is a simple point process where each point is labeled by another random variable, producing a sequence of 2-tuples.

**Definition 2.1.8.** ***Marked Point Process*** [36]: *Let* $\{\mathcal{T}_n \mid n \in N\}$*, where* $N \subset \mathbb{N}$*, be a simple point process defined as above and let* $\{X_n \mid n \in N\}$ *be another stochastic process taking values in* $\mathcal{X} \cup \{\nabla\}$*. Then the sequence of 2-tuples* $\{(\mathcal{T}_n, X_n) \mid n \in N\}$ *is a marked point process if the following two properties are satisfied*

1. $P(X_n \in \mathcal{X}, \mathcal{T}_n < \infty) = P(\mathcal{T}_n < \infty)$
2. $P(X_n = \nabla, \mathcal{T}_n = \infty) = P(\mathcal{T}_n = \infty)$

The nabla denotes values of the mark process, $X$, corresponding to when $\mathcal{T}_n = \infty$; that is, values that will never occur. The marked point process will be important for modeling neurons utilizing a burst multiplexing code, where the mark process will distinguish event spikes from burst spikes.

### Poisson Process on $\mathbb{R}$

The Poisson process with points $\mathcal{T}_n$ is a simple point process satisfying two properties: (1) for any disjoint Borel sets $\{B_i \mid i \in \{1, ..., N\}, N \in \mathbb{N}\}$ the counts $\{\#B_i \mid i \in \{1, ..., N\}, N \in \mathbb{N}\}$, where $\#B_i = \#\{n \mid S_n \in B_i\}$, are mutually independent; (2) $\#B_i \sim \text{poisson}(\lambda)\ \forall i \in \{1, ..., N\}$ [39]. If $\lambda$ is a function on $\mathbb{R}$ then the Poisson process is called inhomogeneous. If $\lambda$ is itself a random variable then the resulting point process is a Cox process. An interesting feature of the Poisson process is that the interval distribution is exponential with rate parameter $\lambda$. The Poisson process is frequently used to model neurons [31].

### Modulated Renewal Process on $\mathbb{R}$

The modulated renewal process generalizes the Poisson process such that the renewal mean rate is a function both of time and of the length of time since the last 'event' (event in this case being a point in the point process rather than an event in the burst-multiplexing sense). Thus, the modulated renewal process intervals will not necessarily be exponentially distributed nor independent of each other. Modulated renewal theory was established by Cox [18, 17] and has been frequently applied to neuroscience [38, 30].

### The Ornstein-Uhlenbeck Process

The final stochastic process we will discuss is the Ornstein-Uhlenbeck (O-U) process, which is frequently used to model noise in the nervous system [20]. Noise, or other signals impinging on a given neuron, typically take the form of point process perturbations of the membrane potential. If one models the decay of the membrane potential to its resting value as exponential this will lead to the stochastic process composed of point process jumps followed by exponential decays, referred to as shot noise [25]. If one chooses two Poisson processes to form the underlying point process, where one leads to positive and the other to negative perturbations, and considers the high rate, low perturbation amplitude limit, the result is an O-U process [68]. This provides motivation to use the O-U process to represent the stochastic component of the temporal evolution of a neuron's membrane voltage. The O-U process is described below.

**Definition 2.1.9.** ***Ornstein-Uhlenbeck process:*** *$\{X_t \mid t \in \mathbb{R}\}$ is a O-U process if it is defined by the following stochastic differential equation*

- $\tau \frac{dX_t}{dt} = -X_t + \sigma dW_t$

*where $W_t$ is the Wiener process (i.e. Brownian motion), $\tau$ is a time constant and $\sigma$ determines the variance of the process.*

The finite dimensional distribution of the O-U process is Gaussian and the process is stationary and of finite memory [24].

### 2.1.3 Dominated Convergence Theorem

One final concept from probability that will be relied upon in this thesis is the dominated convergence, due to Lebesgue [8], which provides a means of exchanging the order of an integral and a limit.

**Theorem 2.1.10.** ***Dominated Convergence Theorem:*** *Let $g$, $f_n$ and $f$ be measurable functions on $\mathbb{R}$ defined w.r.t. probability space $(\Omega, \mathcal{F}, P)$ and assume further that*

1. *$f_n \to f$ as $n \to \infty$ almost everywhere w.r.t. $P$*
2. *$g$ is integrable*
3. *$|f_n| \leq g$ almost everywhere w.r.t. to $P$ and $\forall n$*

*Then $f$ and $f_n$ are also integrable $\forall n$ and $\lim_{n\to\infty} \int_\Omega f_n dP = \int_\Omega \lim_{n\to\infty} f_n dP$*

## 2.2 Network Model

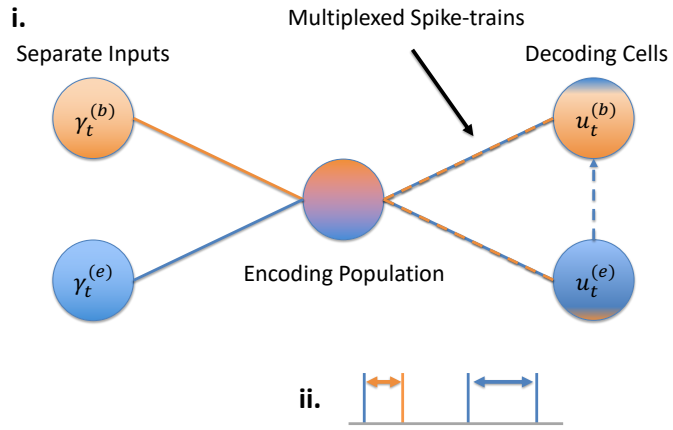


Figure 2.1: **Network Model: i.** Signals $\gamma_t^{(b)}$ and $\gamma_t^{(e)}$ are encoded in the spiking activity of the encoding population (represented by one unit but comprised of many homogeneous cells). The spike-trains produced by these three cells are the multiplexed signals that must then be decoded by the two post-synaptic cells $u_t^{(b)}$ and $u_t^{(e)}$ **ii.** A single multiplexed spike-train is comprised of event spikes (blue) and burst spikes (orange), separated by ISIs with different statistics.

We can distill the problem of neural communication down to how well information can be encoded and subsequently decoded in the brain. Burst multiplexing is a neural population code so this amounts to how well one population of neurons can encode two signals using burst multiplexing, and how well two post-synaptic cells can decode these two signals using STP rules (see section 1.2). We thus need to model the simple biological neural network composed of inputs, encoding population and decoding cells (see figure 2.1). The input signals would be made up of the summed activity of many incoming spike-trains. We will model these inputs using Ornstein-Uhlenbeck processes for the relationship of this class of processes with the Poisson spike-train model (see section 2.1.2). The population of encoding neurons and the decoding cells warrant more detailed descriptions.

### 2.2.1 Encoding Neuron Population

#### Modeling Spike-trains

Before discussing means of generating spike-trains it will be helpful to have a model for a spike-train per se. As a mathematical object, the spike-train is typically convolved with a kernel to obtain either a smoothed spike-rate or an input current for a post-synaptic cell. With this convolutional purpose in mind it makes sense to model a spike-train $S(t)$ as a sum of Dirac delta functions [31]

$$S(t) = \sum_i \delta(t - t^{(i)}) \tag{2.2.1}$$

where $t^{(i)}$ is the time of the $i^{th}$ spike. We can then model the time series representing the activity, $A(t)$, of a population of neurons using a sample mean

$$A(t) = \frac{1}{N} \sum_{j=1}^{N} S_j(t) \tag{2.2.2}$$

where the subscript $j$ denotes the spike-train of the $j^{th}$ neuron in a population of size $N$.

#### Neural Populations

It is expected that the brain encodes a given signal using a population of neurons rather than assigning a single neuron to a single signal. This makes sense intuitively because, to use an unrealistic but illustrative example, if a single neuron encoded the dynamics of the colour red in one's visual field, loss of this neuron would eliminate the ability to perceive red. Thus, a population coding scheme is more robust. Furthermore, to be able to make temporally effective use of any code relying on firing rates, post-synaptic neurons must be able to average over a population of cells to extract an instantaneous rate, rather than having to average over potentially long periods of time. The idea of population coding is supported in the experimental literature [4].

To model a population code identical input signals will be used for every neuron in the encoding population and every neuron will be statistically equivalent– that is, the same model will be used for each. For simplicity we will not model recurrent connections within the encoding population. This is biologically rationalized because there are certain circuits in the brain that are effectively feed-forward [48] and because recurrent effects can be minimalized by balancing them with inhibitory inputs [72]. Thus, modeling the encoding population amounts to determining the model of a single burst multiplexing cell. To place our model we will first provide a brief review of neuron modeling.

### Review of Neuron Models

The first attempts at modeling a neuron were developed from Resistor Capacitor (RC) circuit theory and Ohm's law [12]. Because the voltage difference across the membrane of the cell determines the cell's spiking, one can construct an accurate cell description by modeling the membrane as a capacitor, the voltage dependent ion channels in the membrane as resistors and external inputs as an incoming current. Neglecting spatial effects the dynamics of this RC circuit can be modeled as the voltage at a single point in space via a set of differential equations [35]. Neuron models which only describe a single point in space are referred to as single compartment models. We will discuss two compartment models for burst multiplexing in the following section and, though they exist, multi-compartment models that take into account the spatial extent of a neuron are not relevant for this thesis and will not be developed.

The biological feature requiring more than one dimension to model is the highly nonlinear voltage fluctuation that occurs during a spike. However, the assumption that only the presence or absence of a spike carries information (see section 1.1.2), rather than spike shape, allows the previous model to be simplified to a single differential equation describing sub-threshold dynamics, the dynamics of the neural membrane when no spikes are occurring, and a reset rule: when the sub-threshold dynamics reach a pre-assigned threshold a spike is recorded and the membrane potential is reset to a lower value. This linear differential equation and reset, producing a sub-threshold membrane potential and sequence of spike times, is called the Leaky Integrate and Fire (LIF) model.

Neurons exhibit a memory of the previous spikes they have fired. All neurons exhibit what is called an absolute refractory period, where the neuron cannot fire another spike for a brief time period after spiking. Most neurons also exhibit a relative refractory period, where the neuron is less likely to fire a spike for some short period after spiking. The absolute refractory period can be modeled by incorporating delay in the reset rule of the LIF model and the relative refractory period can be modeled by introducing a second differential equation to describe the memory dynamics. The LIF model with the updated reset rule and second differential equation can sufficiently fit a variety of neural dynamics and is a workhorse of modern computational neuroscience [31].

**Definition 2.2.1.** ***Generalized LIF model:***

- $\tau_v \frac{dv(t)}{dt} = -v(t) + v_0 - \alpha w(t) + \beta I(t)$
- $\tau_w \frac{dw(t)}{dt} = -w(t)$
- *when* $v(t)$ *reaches threshold* $v_{thresh}$ *it is reset to initial condition* $v_0$ *and* $w$ *is increased by the addition of a constant.*

where $I$ is a time dependent input current and $w$ is an adaptation variable dictating spike-memory dynamics, $\tau_v$ and $\tau_w$ are time constants, $\alpha$ and $\beta$ control the relative contributions of spike adaptation and input current to the membrane potential, $v_0$ is the resting membrane potential and $v_{thresh}$ is the threshold at which a spike is emitted.

The models discussed thus far are deterministic; there are two routes that have been taken to introduce stochasticity in neural models. First, one can interchange the differential equations in the above models with stochastic differential equations, by introducing a noisy input usually modeled as white noise or an Ornstein-Uhlenbeck process. Second, one can define a stochastic threshold, where a firing probability is calculated as a function of the membrane potential and spikes are released according to this probability. Members of this second model class are referred to as escape noise models.

Like the LIF, an escape noise model involves a membrane potential and produces a spike-train. The membrane potential is typically framed as an integral equation which depends on the spike history via a refractory kernel $\eta$. The membrane potential is taken as input for a link function whose value determines the firing intensity, or rate function, of the point process generating the spike-train. This model is called the Spike Response Model (SRM)

**Definition 2.2.2.** ***$SRM_0$ and SRM:***

- $v(t) = \eta\left(t - t'\right) + \int_0^\infty \kappa(s) I(t-s) ds + v_0$, *where* $t'$ *is the time of the last spike*
- $\lambda(t) = f_{\text{link}}\big(v(t)\big)$
- $S_t$ *is generated by a renewal point process with rate* $\lambda(t)$
- *If the first term in the definition of* $v(t)$ *is replaced by a sum over the refractory kernels associated with all previous spikes,* $\sum_i \eta\left(t - t^{(i)}\right)$, *then the model is referred to as the* $SRM$. *Note that the* $SRM_0$, *but not the* $SRM$, *is a renewal process.*

where $\kappa$ is a kernel producing a voltage from the input current, $S_t$ is the generated spike-train, now a stochastic process, and everything else is as previously defined. The refractory kernel, $\eta$, and input kernel, $\kappa$, are typically taken to be exponential functions as the sub-threshold membrane potential that these aim to model is well described by linear differential equations; the kernel parameters are then estimated from experimental data [31]. The link function $f_{\text{link}}$ can be modeled as a rectified linear function or exponential as its primary purpose is to map membrane potential to a non-negative firing rate. Finally, note that both $v$ and $\lambda$ depend implicitly on the spike history and that the $SRM$ models define point processes.

### Studied Neuron Model

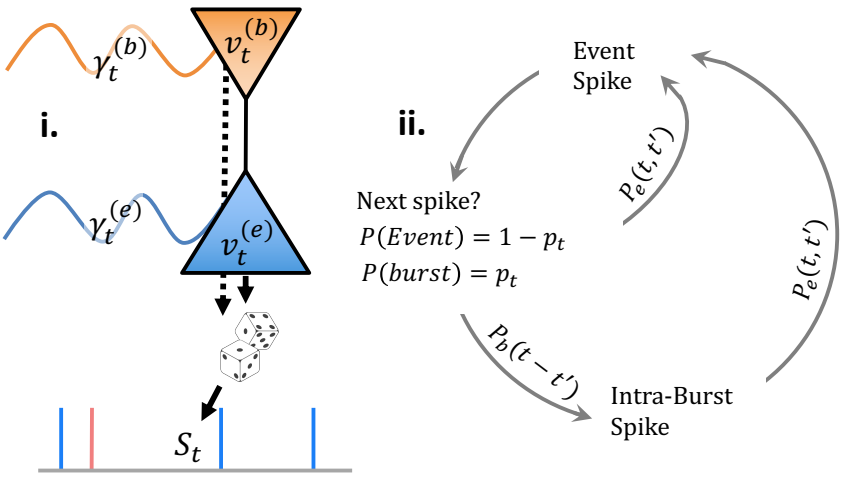


Figure 2.2: ***SRM*$_0^2$ Model: i. Model Schematic.** Signals $\gamma_t^{(b)}$ and $\gamma_t^{(e)}$ are inputs to the burst and event compartments of the the model respectively, making the membrane potentials, $v_t^{(b)}$ and $v_t^{(e)}$, random variables also. These two membrane potentials generate a spike-train, $S_t$, via a marked point process. **ii. Model Flow Chart.** Beginning at the top: an event spike occurs and a Bernoulli random variable is sampled to determine the identity of the next spike. If a burst occurs a gamma distribution is sampled and the burst spike is added based on the sum of this gamma random variable and the spike absolute refractory period. If an event spike occurs the event rate dictates the occurrence of the next spike. Only an event spike can occur after a burst spike; short absolute refractory periods where the rate is zero occur after every spike.

Burst multiplexing adds a level of complexity to the standard definition of the spike-train because now one must consider the identity of the spike, whether the spike is an event spike or a burst spike. To our knowledge there are only two burst-multiplexing models that have currently been published. The first is a two compartment version of the LIF model, where one compartment is used to integrate information for burst generation and the other is used for event generation. This is

designed in an effort to model the layer 5 pyramidal cells discussed in section 1.2.2. The second model is a marked $\text{SRM}_0$ that does not make explicit the intra-burst ISIs, rather a single event in the point process is marked as a burst or an event and the burst spike itself (see section 1.2 for definitions of *burst*, *event*, *burst spike* and *event spike*), when it occurs, is assumed to fall in the absolute refractory period of the event spikes and is thus not modeled. Both of these models were published in Naud & Sprekeler's work [54].

Our model consists of a minor extension to the marked $\text{SRM}_0$ model employed by Naud & Sprekeler. Our research question requires computational efficiency to generate the large data sets necessary for calculating information rate but also requires explicit representation of intra-burst interval to relate information to ISI distribution. The marked $\text{SRM}_0$ option was chosen over the LIF model because it represented a natural extension of Naud & Sprekeler's work and because we expect that it is more computationally efficient than the LIF, though we did not perform any benchmarking. We updated the $\text{SRM}_0$ to model both burst spikes and event spikes, however we assume bursts that contain at most one burst spike (bursts that are only two spikes long). Because of the two compartment nature of our model we refer to it as $\text{SRM}_0^2$ (see figure 2.2).

**Definition 2.2.3.** $\boldsymbol{SRM_0^2}$ **:** *is a self-inhibiting marked point process defined by the double sequence* $\{(B_n, \mathcal{T}_n) \mid n \in \mathbb{N}\}$ *where* $\mathcal{T}_n$ *is a process with rate* $\lambda_t$ *and is constructed by alternating, as a function of the mark sequence, between sampling a modulated renewal process and a renewal process. The rate is thus dependent on the mark process and is defined as follows*

- $v^{(b)}(t) = v_{\text{in}}^{(b)}(t) + v_0$ *burst generating compartment membrane potential*
- $v^{(e)}(t) = \eta(t-t') + v_{\text{in}}^{(e)}(t) + v_0$ *event generating compartment membrane potential*
- $p(t) = f_{\text{link,b}}\big(v^{(b)}(t)\big) \in [0,1]$ *burst probability*
- $\rho(t) = f_{\text{link,e}}\big(v^{(e)}(t)\big) \in [0,\infty)$ *event rate in absence of bursts*
- $\Gamma(t)$ *is the rate function for sampling an interval that is Gamma distributed with scale parameter* $\Gamma_1$ *and shape parameter* $\Gamma_2$
- $B_n = X \sim \text{Bernoulli}\big(p(\mathcal{T}_n)\big)$
- $\lambda_t = \begin{cases} \rho(t) & \text{if} \quad t \in (\mathcal{T}_n, \mathcal{T}_{n+1}) \text{ and } B_n = 0 \\ \Gamma(t - \mathcal{T}_n) & \text{if} \quad t \in (\mathcal{T}_n, \mathcal{T}_{n+1}) \text{ and } B_n = 1 \\ 0 & \text{if} \quad t - \mathcal{T}_n < \Delta_{\text{ref}} \end{cases}$
- $S_t = \sum_n \delta(t - \mathcal{T}_n)$ *is the generated spike train.*

where $v_{\text{in}}^{(b)}(t)$ and $v_{\text{in}}^{(e)}(t)$ are time-varying inputs. In this model the marked process labels the first spike in a burst with a 1 and all other spikes with 0. In this way the set of burst spikes includes every spike following a 1-marked spike and the set of event spikes is the complement of the burst spike set. In analogy to the link function of the $\text{SRM}_0$, the purpose of the link functions in the $\text{SRM}_0^2$ is to map membrane potential to image spaces appropriate for probability and rate respectively, and thus were taken to be a sigmoid for the burst compartment and an exponential for event. For full parameterization of the model see Appendix A.1. As the purpose of this paper is to explore the problem of information in a burst modeling network, rather than engaging in a detailed study of point processes, we have not performed any rigorous analysis of the $\text{SRM}_0^2$ model. We simply assume that it generates spike train processes appropriate for the goals of our study, a phenomenon that has been verified numerically.

### 2.2.2 Decoding Cells

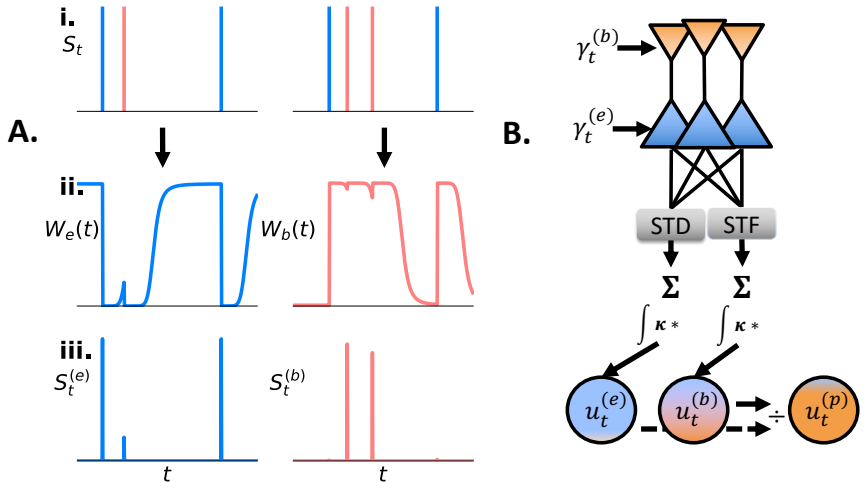


Figure 2.3: **Weight Functions & Model Schematic A. Weight Functions.** STP is modeled using weight functions: multiplying $S_t$ with a weight function selects for only spikes with a specific ISI interval. **i.** two spike-trains. **ii.** weight functions selecting events (left) and bursts (right) from above spike-trains. **iii.** Multiplying original spike-trains with weight functions produces approximate event (left) and burst (right) trains. **B. Model Schematic.** Beginning at the top: signals are encoded in population; spike-trains are decoded by STD and STF, represented by weight functions, on a cell-by-cell basis; weighted spike-trains are summed and convolved to produce cell potentials estimating event, $u_t^{(e)}$, and burst, $u_t^{(b)}$, rates; quotient of these is taken to approximate third cell potential, $u_t^{(p)}$, estimating burst fraction.

Under a burst multiplexing hypothesis the goal of the post-synaptic cells in a neural circuit is to take spike-trains from the pre-synaptic population, decode $\gamma_1$ and $\gamma_2$ from them using STP rules and then average the decoded signals over the pre-synaptic population to obtain better estimates. The key biological features that must be modeled here are the STP rules and, in particular, the errors brought about by using these imperfect decoding rules–STD labelling a burst spike as an event or STF labeling an event spike as a burst.

We model the STP rules using weight functions that multiply each neuron's spike-train individually to scale the response of the post-synaptic neurons to every pre-synaptic spike as a function of the ISIs that came before it. Because we are only modeling bursts of two spikes there is no extra information about the spike identity

(burst or event) in the spike-train beyond the ISI directly before the spike. As such, we will use weight functions that depend only on the distance between the previous spike time, $t'$, and the current time, $t$. Specifically, we choose $w_e(t-t') = \Theta(t-t'-\tau_{\mathrm{w},e})$ for the event weight function and $w_b(t-t') = \Theta(t'-t+\tau_{\mathrm{w},b})$ for the burst weight function, where $\Theta$ is the Heaviside and $\tau_{\mathrm{w},i} \mid i \in \{e,b\}$ is a threshold parameter. Intuitively what this means is that we have a hard threshold such that all the spikes that occur at the end of an ISI greater than $\tau_{\mathrm{w},e}$ are considered events and all spikes that occur at the end of an ISI less than $\tau_{\mathrm{w},b}$ are considered bursts. The rational for using Heaviside weight functions is that they capture the general behaviour of the STP rules and the issue of spike misclassification while relying on a single, easily interpretable parameter. This parameter was then selected to maximize information transmission from a set of biologically reasonable values.

After applying the weight functions to each spike-train we sum the result and convolve the sum with a kernel representing the effect of the incoming spikes on the membrane potential of the post-synaptic cell. For the post-synaptic cell using the event weight function this results in an estimate of the event rate, yielding one of the two outputs of our model system. There is one more step that remains however for estimating burst fraction. The membrane of the cell using the burst weight function to decode spikes is estimating the burst rate, which is not an estimate of the encoded burst generating signal because it is still heavily corrupted by the event signal: when the event rate goes up more bursts are produced and when the event rate goes down less bursts are fired, even if the burst signal is constant. To extract the burst fraction we must divide the burst rate by the event rate, an operation that has been shown to be implementable by neural machinery (i.e. divisive inhibition [54, 23]). Rather than explicitly modeling such neural machinery we simply take the ratio of the burst and event rates. We summarize our decoding model below

**Definition 2.2.4.** ***Decoding Machinery:***

1. $S_t^{(b)} = \sum_i w_b(t-t^{(i)})\delta(t-t^{(i)}), \quad S_t^{(e)} = \sum_i w_e(t-t^{(i)})\delta(t-t^{(i)})$
2. $A_t^{(b)} = \frac{1}{N}\sum_{j=1}^{N} S_t^{(b,j)}, \quad A_t^{(e)} = \frac{1}{N}\sum_{j=1}^{N} S_t^{(e,j)}$, *$N$ is the encoding population size*
3. $u_t^{(b)} = N[\kappa_{syn} * A^{(b)}](t) + v_{syn}, \quad u_t^{(e)} = N[\kappa_{syn} * A^{(e)}](t) + v_{syn}$
4. $u_t^{(p)} = u_t^{(b)}/u_t^{(e)}$

*1. the decoded event, burst spike trains; 2. the decoded event, burst population activity; 3. the membrane filtered event, burst activities. The former is an estimate of the event generating signal; 4. the estimated burst generating signal.*

### 2.2.3 Model Summary

In summary, our model takes two Ornstein-Uhlenbeck processes as inputs, one to modulate burst fraction and one to modulate event rate. These inputs are encoded in the spike-trains of a population of $N$ marked renewal process neurons. Each spike-train is decoded using a weighting function modeling STD and the total $N$ spike-trains are summed and convolved with a synaptic filter to produce an estimate of the input event-related input signal. Each of the spike-trains is also decoded using a weight function modeling STF to arrive at an estimate of burst rate. The quotient of the estimated burst and event rates is then taken to estimate burst fraction. See figure 2.3 for a schematic of the full model.

## 2.3 Signal Processing

We now turn to the issue of analyzing the network model. To this end a numerical approach was adopted where we simulate samples from the model and then apply information theoretic techniques to the generated data. As is the case in many fields, it often makes sense to map data that one wishes to analyze into a new representation space in which it is easier to perform the analysis. A natural candidate for the representation space when dealing with time series is the frequency domain, the space whose coordinates are frequencies rather than points in time. It will become clear in this section that representing stationary stochastic processes in the frequency domain carries statistical benefits over time domain representations. To begin with, we will review the Fourier transform, complex random variables and the spectral density. We will then explain the benefits of the frequency domain representation, which follow, in the discrete-time case we ultimately consider, from results due to Brillinger.

### 2.3.1 Fourier Transform

A periodic function $x : \mathbb{R} \to \mathbb{R}$ that satisfies certain regularity and smoothness conditions can be represented by a Fourier series, the limit of a sum of complex exponentials of varying frequency and phase [49].

**Definition 2.3.1.** ***Fourier Series representation of*** $x(t) \in \mathbb{R}$ **:**

$$x(t) = \lim_{T\to\infty} \sum_{n=-\infty}^{\infty} \hat{x}^{(T)}\Big(\frac{n}{2T}\Big) \exp\Big(\frac{\pi i n t}{T}\Big) \frac{1}{\sqrt{2T}} \tag{2.3.1}$$

*where* $T \in \mathbb{R}$*,* $i$ *is the imaginary number and the Fourier series coefficients are defined thus*

$$\hat{x}^{(T)}(f) = \frac{1}{\sqrt{2T}} \int_{-T}^{T} x(t) \exp\left(-2\pi i f t\right) dt \tag{2.3.2}$$

Before taking the limit $T \to \infty$ the function will have period $2T$. The type of convergence exhibited by the limit in equation 2.3.1 depends on the conditions that $x$ satisfies; for example, if $x$ is absolutely integrable and of bounded variation on one period then we have point-wise convergence [11]. However, these issues of existence disappear when one widens one's scope to generalized functions, as every generalized function has a Fourier transform given by another generalized function [49]. If the restrictions on $x$ are appropriate the series in equation 2.3.1 converges to an integral in the limit, thus allowing for an aperiodic (period $2T = \infty$) function to be represented as an integral over frequency components.

Through Euler's formula, the Fourier series can be written as a sum of cosine and sine functions. This representation provides intuition into what the Fourier transform at $f$ is doing: quantifying how much the sine and cosine functions at frequency $f$ are present in the original signal, via the amplitude (modulus) of the complex exponential, and what the relative proportion of the sine and cosine functions at this frequency are, via the phase (argument) of the complex exponential.

If one wishes to perform data analysis on the frequency domain representation of a continuous time signal one must of course map their signal to the frequency domain. This is done using the $\lim_{T\to\infty}$ of equation 2.3.2 to obtain the frequency values $\hat{x}(f)$. There are two obvious issues here, however. First, it is impossible to obtain an infinitely long sample of a time series. Second, it is equally impossible to sample a time signal continuously, and analysis algorithms are implemented digitally so one can only manipulate discrete signals. For these two reasons it is necessary to approximate the infinite $T$ limit of equation 2.3.2. This is done by first choosing a large enough value of $T$ that $\hat{x}(f) \approx \hat{x}^{(T)}(f)$. Second, the integral in equation 2.3.2 is approximated using a Riemann sum

$$\hat{x}^{(T)}(f) \approx \frac{1}{\sqrt{2T}} \sum_{n=-N}^{N} x(n\Delta t) \exp\left(-2\pi i f n \Delta t\right) \Delta t \tag{2.3.3}$$

where we define $\Delta t = \frac{T}{N}$ s.t. $N \in \mathbb{N}$. The discrete sampling of the time domain means that the frequency domain must also be sampled discretely. It becomes impossible to estimate frequencies at a higher resolution than $\frac{1}{2T}$ or with periods shorter than $2\Delta t$. More formally, this means that an arbitrary frequency component must be approximated as follows: $\hat{x}(f) \approx \hat{x}(k\Delta f)$ where $\Delta f = \frac{f_{\max}}{N}$ s.t. $f_{\max} = \frac{1}{2\Delta t}$ and $k \in \{-N, -(N-1), ..., N-1, N\}$. Taking this final approximation into account yields the following

$$\hat{x}(f) \approx \sqrt{\frac{\Delta t}{2N}} \sum_{n=-N}^{N} x_n \exp\left(\frac{-2\pi i n k}{2N}\right) \tag{2.3.4}$$

where we have let $x_n = x(n\Delta t)$. The RHS of equation 2.3.4 is the definition of the Discrete Fourier Transform (DFT) multiplied by constant $\sqrt{\Delta t}$. The DFT is a discrete analog to the continuous Fourier transform and is easily computable using the Fast Fourier Transform algorithm [74]. The above shows that the DFT can be used to approximate the frequency components of a continuous signal.

Importantly, a foundational theorem in signal processing–the Nyquist-Shannon Theorem [66]–states that bandwidth limited continuous signals, signals whose frequency representations are zero above a certain threshold, can be completely described by a discrete sampling when the sample rate is sufficiently high. This means

that, for bandwidth limited signals, the only approximation being made when dealing with the DFT comes from approximating an infinitely long signal with a finite sample of the signal.

Three properties of the Fourier transform will be important for this thesis: orthogonality, symmetry for real inputs and shift [10]. The DFT is obviously linear and can thus be formulated as a matrix. Using the $\frac{1}{\sqrt{N}}$ normalization, as we have, the DFT is an orthogonal transformation so the matrix DFT is unitary. When the Fourier transform, discrete or continuous, is applied to real signals the resultant Fourier components are symmetric in the sense that $\hat{x}(f) = \overline{\hat{x}(-f)}$. Furthermore, the Fourier transform of a function, $x$, satisfying $x(t) = x(-t)$ is real. Finally, one can choose the index of the DFT to be $\{0, 1, ..., 2N-1, 2N\}$, rather than $n \in \{-N, -(N-1), ..., N-1, N\}$ (see section C.2), and the resultant Fourier components will only differ by a constant factor of magnitude 1. Thus, formulations using either indices are essentially equivalent and we will switch back and forth between the two, depending on circumstances, from this point onwards.

### 2.3.2 Spectral Densities

The power spectrum of a stochastic process can be thought of as the frequency representation of the temporal correlation structure of the process and, accordingly, is defined as the un-normalized Fourier transform of the autocorrelation function associated with the process [56]

**Definition 2.3.2.** ***Power Spectrum:*** *If $X$ is a stationary stochastic process then its power spectrum, defined as a function of frequency, is given by*

$$\mathcal{P}_{XX}(f) = \sum_{\tau} R_{XX}(\tau) \exp(-2\pi i f \tau)$$

*where $f \in \mathbb{R}$ and the sum is over all values of $\tau \in \mathbb{Z}$.*

The autocorrelation function is symmetric so the power spectrum is real-valued. The cross-spectrum generalizes the power spectrum analogously to the relationship between autocorrelation and cross-correlation:

**Definition 2.3.3.** ***Cross Spectrum:*** *If $X$ and $Y$ are stationary stochastic processes then their cross spectrum, defined as a function of frequency, is given by*

$$\mathcal{P}_{XY}(f) = \sum_{\tau} R_{XY}(\tau) \exp(-2\pi i f \tau)$$

*where $f \in \mathbb{R}$ and the sum is over all values of $\tau \in \mathbb{Z}$.*

### 2.3.3 Complex Random Variables

Complex random variables come about naturally when one applies the discrete Fourier transform to a vector of random variables and, as such, are important for this thesis. As the name suggests, a complex random variable is defined in the same fashion as a random variable except instead of mapping from a probability space to $\mathbb{R}$ it is a mapping to $\mathbb{C}$. Because of the isomorphism between $\mathbb{R}^2$ and $\mathbb{C}$ one can equivalently think of a complex random variable as a 2-dimensional random vector. Unlike with random variables, there is not a single $k^{th}$ moment for a given complex random variable but several [26]. If $Z$ is a complex random variable this is seen for its $2^{nd}$ central moment in the necessity for two moment-related values: the variance $E\big([Z - E(Z)][\overline{Z - E(Z)}]\big)$ and the pseudo-variance $E\big([Z - E(Z)][Z - E(Z)]\big)$. A particular type of complex normal random variable will play a key role later in this section so we will discuss it briefly below.

#### Circularly Symmetric Complex Normal Distribution

The circularly symmetric complex normal distribution is a particular subtype of multivariate, normally distributed, complex-valued random vector. Note that we will sometimes refer to it simply as the complex normal distribution in the following pages. It is defined as follows

**Definition 2.3.4.** ***Complex Normal Random Vector*** [11]: *Let* $X = \mathrm{Im}(X) + i\mathrm{Re}(X)$, $\mathrm{Im}(X), \mathrm{Re}(X) \in \mathbb{R}^n$, *and* $Z = [\mathrm{Re}(X) \quad \mathrm{Im}(X)]^T$. *We say* $X \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \Sigma)$, *where* $\Sigma_{ij} = \mathrm{cov}(X_i; X_j)$, *if* $Z$ *is normally distributed with*

- *mean:* $[\mathrm{Re}(\boldsymbol{\mu}) \quad \mathrm{Im}(\boldsymbol{\mu})]^T$
- *covariance matrix:* $\frac{1}{2}\begin{bmatrix} \mathrm{Re}(\Sigma) & -\mathrm{Im}(\Sigma) \\ \mathrm{Im}(\Sigma) & \mathrm{Re}(\Sigma) \end{bmatrix}$

Defined thus, the complex normal distribution has two nice properties. First the pseudo-covariance matrix is the zero matrix, thus requiring only mean vector and covariance matrix to describe the complex normal random vector. Second, the real and imaginary components of $X$ are independent if $\Sigma$ is diagonal.

### 2.3.4 Brillinger's Theorem

Brillinger's theorem provides a kind of central limit theorem for the discrete Fourier transform, giving strong motivation for working in the frequency domain when analyzing long time series. We provide the proof of the theorem here both because the result is central to this thesis and, as the original statement of the proof is subdivided into several theorems and lemmas, we believe having the components of the proof collected and presented together is useful.

**Theorem 2.3.5.** ***Brillinger's Theorem*** [11]: *Let $X$ be a stochastic process taking values in $\mathbb{R}^n$. If $X$ satisfies*

- *Finite memory*
- *Jointly stationary element processes*
- $\mathbb{E}(X_t) = \vec{0}$

*Then $\hat{X}_T(f) \xrightarrow{d} \hat{X}_f$ as $T \to \infty$ where $\hat{X}_f \sim \mathcal{N}_{\mathbb{C}}(\vec{0}, \Sigma)$ such that the elements of $\hat{X} = \{\hat{X}_f \mid f \in [0, \frac{1}{2\Delta t}]\}$ are mutually independent and $\Sigma_{ij} = \mathcal{P}_{X^{(i)}X^{(j)}}(f)$ is the cross spectrum of the $i^{th}$ and $j^{th}$ element processes. $\Delta t$ is the distance in the index space between consecutive values of the stochastic process (e.g., the sample rate if $X$ is formed by discretely sampling a continuous time process).*

**Proof:** Let $\left\{X(t)\right\}_{t=-\infty}^{\infty}$ $t \in \mathbb{Z}$, $X(t) \in \mathbb{R}^n$ $\forall t$ be a stochastic process such that $\left\{X^{(l)}(t)\right\}_{t=-\infty}^{\infty}$ $l \in \{1, ..., n\}$ is the stochastic process at the $l^{th}$ element of $X$. Assume that $X$ is a strictly stationary process (thus its elements are jointly stationary) with finite memory. We also define the window function $h_l : \mathbb{R} \to \mathbb{R}$ s.t. $h_l(x) = 0$ $\forall x \notin [-T+1, T-1]$. Thus $h_l(x)$ implicitly depends on $T$. We assume that $h_l$ is of bounded variation $\forall\, l$ and $T$ and we adopt the following notations to simplify what is to come:

$$\mathcal{H}_T(f) \coloneqq \sum_{t=-T+1}^{T-1} \prod_{j=1}^{k} h_{l_j}(t) e^{-2\pi i f t} \tag{2.3.5}$$

$$\hat{X}_T^{(l)}(f) = \frac{1}{\sqrt{2T-1}} \sum_{t=-T+1}^{T-1} X^{(l)}(t) h_l(t) e^{-2\pi i f t} \tag{2.3.6}$$

Note that the formulation of the DFT we are adopting here is as in equation 2.3.3 except we have chosen the time step, $\Delta t$, to be 1 for simplicity (the proof would proceed via the same steps if the time step was different). Because of this choice $T = N$ and we will thus use $T$ in the fashion that $N$ has been previously used, by defining $T \in \mathbb{N}$. Finally, we will also use our previously defined (see section 2.1.1) notation for the cumulants of stochastic processes. We now consider the cumulants of the finite Fourier transforms of the element processes of $X$. Using the linearity of the cumulant we can write this as

$$\mathcal{C}\left[\hat{X}_T^{(l_1)}(f_1), ..., \hat{X}_T^{(l_k)}(f_k)\right] = \frac{1}{(2T-1)^{\frac{k}{2}}} \sum_{t_1,...,t_k=-T+1}^{T-1} \prod_{j=1}^{k} h_{l_j}(t_j) \mathcal{C}_{\mathcal{S}_k}(t_1, ..., t_k) e^{-2\pi i \sum_{j=1}^{k} f_j t_j} \tag{2.3.7}$$

using the stationarity assumption this is equivalent to

$$= \frac{1}{(2T-1)^{\frac{k}{2}}} \sum_{t_1,...,t_{k-1}=-T+1}^{T-1} \sum_{t_k=-T+1}^{T-1} \prod_{j=1}^{k} h_{l_j}(t_j) \mathcal{C}_{\mathcal{S}_k}(t_1 - t_k, ..., t_{k-1} - t_k, 0) e^{-2\pi i \sum_{j=1}^{k} f_j t_j} \tag{2.3.8}$$

$$= \frac{1}{(2T-1)^{\frac{k}{2}}} \sum_{t=-T+1}^{T-1} \sum_{u_1,...,u_{k-1}=-T+1-t}^{T-1-t} \prod_{j=1}^{k-1} h_{l_j}(u_j + t) h_{l_k}(t) \mathcal{C}_{\mathcal{S}_k}(u_1, ..., u_{k-1}, 0) e^{-2\pi i \sum_{j=1}^{k-1} f_j u_j} e^{-2\pi i \sum_{j=1}^{k} f_j t} \tag{2.3.9}$$

where for the second equality we have made the substitutions $u_j = t_j - t$ and $t_k = t$. Note that, given the definition of $h_l$, we can extend the boundaries of the second summation to $\pm S = \pm 2(T-1)$ without changing the sum's value, thus making the cumulants independent of $t$. Using this observation and rearranging yields

$$= \frac{1}{(2T-1)^{\frac{k}{2}}} \sum_{u_1,...,u_{k-1}=-S}^{S} \mathcal{C}_{\mathcal{S}_k}(u_1, ..., u_{k-1}, 0) e^{-2\pi i \sum_{j=1}^{k-1} f_j u_j} \sum_{t=-T+1}^{T-1} \prod_{j=1}^{k-1} h_{l_j}(u_j + t) h_{l_k}(t) e^{-2\pi i \sum_{j=1}^{k} f_j t} \tag{2.3.10}$$

Before proceeding further, we note that the below difference is bounded in the following fashion, where $K$ is a constant

$$\sum_{t=-T+1}^{T-1} \prod_{j=1}^{k-1} h_{l_j}(u_j + t) h_{l_k}(t) e^{-2\pi i \sum_{j=1}^{k} f_j t} - \mathcal{H}_T\Big(\sum_{j=1}^{k} f_j\Big) \leq K \sum_{j=1}^{k-1} |u_j| \tag{2.3.11}$$

This follows from $h_l$ having bounded variation and from definition 2.3.5. Adding and subtracting the second term on the LHS in (2.3.11) appropriately in (2.3.10) yields (2.3.12)

$$(2.3.10) = \frac{\mathcal{H}_T\Big(\sum_{j=1}^{k} f_j\Big)}{(2T-1)^{\frac{k}{2}}} \sum_{u_1,...,u_{k-1}=-S}^{S} \mathcal{C}_{\mathcal{S}_k}(u_1, ..., u_{k-1}, 0) e^{-2\pi i \sum_{j=1}^{k-1} f_j u_j} + \epsilon_T \tag{2.3.12}$$

where the $\epsilon_T$ term is bounded as shown in the two lines of equations below, using (2.3.11).

$$|\epsilon_T| \leq K \sum_{u_1,...,u_{k-1}=-S}^{S} \frac{\sum_{j=1}^{k-1} |u_j|}{(2T-1)^{\frac{k}{2}}} \left| \mathcal{C}_{\mathcal{S}_k}(u_1, ..., u_{k-1}, 0) \right|$$

$$= K \sum_{u_1,...,u_{k-1}=-\infty}^{\infty} \frac{\sum_{j=1}^{k-1} |u_j|}{(2T-1)^{\frac{k}{2}}} \left| \mathcal{C}_{\mathcal{S}_k}(u_1, ..., u_{k-1}, 0) \right| \mathbb{1}_{u_j \in [-S,S], j \in [1,...,k-1]}(\mathbf{u}) \tag{2.3.13}$$

Here we have made the definition $\mathbf{u} \coloneqq [u_1 ... u_{k-1}]^T$.

We are interested in the limiting distribution of the DFT so we will now take the limit $T \to \infty$ in (2.3.12). Consider the sum over all $u_j$ in (2.3.13) to be a Lebesgue integral w.r.t. the counting measure on $\mathbb{Z}^{k-1}$. By $\frac{\sum_{j=1}^{k-1} |u_j|}{(2T-1)^{\frac{k}{2}}} \leq 1$ for $k \geq 2$ and the finite memory of $X$ we can apply dominated convergence theorem to take the limit $T \to \infty$ inside the integral (infinite sum), which implies $\epsilon_T \to 0$ as $T \to \infty$. This means that all $k^{th}$ order cumulants of the DFT converge to the first term in (2.3.12).

Let us now select a window function consistent with the finite DFT

$$h_l(x) = \begin{cases} 1 & x \in [-T+1, T-1] \\ 0 & \text{else} \end{cases} \tag{2.3.14}$$

Using this definition in 2.3.12 yields

$$\frac{\sum_{t=-T+1}^{T-1} e^{-2\pi i \left( \sum_{j=1}^{k} f_j \right) t}}{(2T-1)^{\frac{k}{2}}} \sum_{u_1,...,u_{k-1}=-S}^{S} \mathcal{C}_{\mathcal{S}_k}(u_1, ..., u_{k-1}, 0) e^{-2\pi i \sum_{j=1}^{k-1} f_j u_j} + \epsilon_T \tag{2.3.15}$$

It is clear that the numerator of the first fraction in this expression is bounded by $2T$ which, together with finite memory and the limiting behaviour of $\epsilon_T$, implies that all cumulants of order 3 and above for $\hat{X}_T(\boldsymbol{f})$, where $\boldsymbol{f}$ is the vector of $k$ frequencies, converge to zero in the limit of large $T$, implying that the Fourier components are circularly symmetric complex normal.

We will now evaluate the cases $k = 1$ and $k = 2$ respectively. The $k = 1$ case is the expected value of the given Fourier component. By assumption $E(X_t) = 0$ so it trivially follows that $E(\hat{X}_f) = 0 \, \forall \, f$. Note that this assumption is a minor one as any stationary process can be centered at 0. Consider now the case where $k = 2$. Specifically, we are interested in $\operatorname{cov}\left[\hat{X}_T^{(a)}(f_T^{(a)}), \hat{X}_T^{(b)}(f_T^{(b)})\right]$, where we continue to use the subscript $T$ to denote dependence on $T$. Because the mean is equal to zero the covariance is simply equal to $E\left[\hat{X}_T^{(a)}(f_T^{(a)}), \overline{\hat{X}_T^{(b)}(f_T^{(b)})}\right]$. Because of the conjugation

here the second frequency will be negative, allowing us to write the first term in equation 2.3.15 as

$$\mathcal{H}_T\Big(\sum_{j=1}^{k} f_j t\Big) = \sum_{t=-T+1}^{T-1} e^{-2\pi i[f_T^{(a)} - f_T^{(b)}]t} \tag{2.3.16}$$

Let $f_T^{(i)} = r_i \Delta f$ for $i \in \{a, b\}$, $r_i \in \mathbb{Z}$, as in section 2.3.1, and assume $f_T \to f$ as $T \to \infty$. Then

$$\sum_{t=-T+1}^{T-1} e^{-2\pi i[f_T^{(a)} - f_T^{(b)}]t} = \sum_{t=-T+1}^{T-1} e^{-2\pi i \frac{r_{(a)} - r_{(b)}}{2T-1} t} \tag{2.3.17}$$

From the above expression it can be shown that (see Appendix C.1 for details)

$$\mathcal{H}_T = \begin{cases} \mathcal{O}(2T-1) & \text{if} \quad r_a \pm r_b \equiv 0 (\text{mod} 2T-1) \\ \mathcal{O}(1) & \text{else} \end{cases} \tag{2.3.18}$$

If $f_T^{(i)} \in [0, \frac{1}{2\Delta t}]$ then $|r_i| \leq T-1$ and thus the Fourier components at different frequencies are asymptotically independent on $[0, \frac{1}{2\Delta t}]$. Mutual independence of different frequencies follows from the fact that the Fourier components are normally distributed.
Now consider the covariance matrix, $\Sigma$, of the random vector $\hat{X}_f$ composed of the Fourier components at frequency $f$ of the element processes of $X$. Equations 2.3.15 and 2.3.18 give

$$\Sigma_{ab} = \lim_{T\to\infty} \text{cov}\big[\hat{X}_T^{(a)}(f_T), \hat{X}_T^{(b)}(f_T)\big]$$

$$= \lim_{T\to\infty} \sum_{u=-2(T-1)}^{2(T-1)} \mathcal{C}_{\mathcal{S}_2}(u, 0) e^{-2\pi i f_T u} = \mathcal{P}_{X^{(a)}X^{(b)}}(f) \tag{2.3.19}$$

∎

## 2.4 Information Theory for Stochastic Processes

Recall that the goal of this thesis is to measure how well information is transmitted under a burst-multiplexing code and various spike train statistics. To do so we must find some means of quantifying information transmission between the

inputs to the encoding neurons in the model and post-synaptic, decoded, signals; we must quantify information between stochastic processes. In the next section 2.5 we will investigate methods for analyzing information transmission between stochastic processes which we will then apply to our neural network model in the results chapter (section 3.2). The ideas that underly these methods come from Shannon's information theory so this section reviews information theoretic concepts, entropy and mutual information, then describes extensions of these ideas to stochastic processes, entropy and information *rates*, in preparation for the next section.

### 2.4.1 Entropy and Information

The topics of this section, entropy and mutual information, come from information theory, a field established at the end of World War II by Claude Shannon to provide a framework for understanding problems in communication [65].

#### Entropy

Entropy is a property of a random variable, or vector. Though entropy is, in its original formulation, defined for discrete random variables an analogous quantity known as differential entropy extends the idea of entropy to continuous variables. During this thesis we will often refer to both entropy and differential entropy simply as "entropy" and we will adopt the same symbol for both. The class of random variable under consideration will make it obvious which of the two is being discussed. While the interest of this project is primarily in differential entropy, entropy for discrete random variables will also be reviewed because it provides a nice perspective on the concept more generally

**Definition 2.4.1.** ***Entropy and Differential Entropy:*** *Let $X$ be a discrete (continuous) random variable or vector with probability mass function (density) $p$. Then its (differential) entropy (when it exists) is defined*

$$H(X) = -E\big(\log_2 p(x)\big) \tag{2.4.1}$$

It follows that one can consider joint entropies in either the continuous or discrete case.

**Definition 2.4.2.** ***Joint Entropy:*** *Let $X$ and $Y$ be discrete (continuous) random variables with joint mass function (density) given by $p(x, y)$. Then their joint entropy (differential entropy) is*

$$H(X, Y) = -E\big(\log_2 p(x, y)\big) \tag{2.4.2}$$

Entropy is non-negative while differential entropy takes values on $\mathbb{R}$. Intuitively, entropy provides a measure of the disorder of a random variable. For example, the probability mass function with all its mass on a single value contains no disorder in the sense that there is no uncertainty in which value it will exhibit. Accordingly, the degenerate random variable associated with this mass function will have 0 entropy. In the same vain, if $X$ and $Y$ are Gaussian random variables but $V(X) > V(Y)$ then $H(X) > H(Y)$ because there is more disorder, or uncertainty, in the higher variance case.

The $\log_2$ base is typically used in the formulation of entropy because it provides values in units of bits. This gives another perspective on the meaning of entropy, that it is the mean number of bits necessary to describe a sample from the random variable. For example, if $X$ is Bernoulli with parameter $p = 0.5$ one requires on average 1 bit to describe a sample (e.g. 0 for $X = 0$ and 1 for $X = 1$). This intuition applies universally in the case of discrete random variables but doesn't fully translate to the continuous case because a negative mean number of bits doesn't make sense physically.

The following lemma regarding differential entropy will be useful later on. We assume in all cases that the differential entropies are well defined.

**Lemma 2.4.3.** ***Differential Entropy of the Multivariate Gaussian*** [16]: *If $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and takes values in $\mathbb{R}^n$. Then*

$$H(X) = \frac{1}{2} \log_2 \big( (2\pi e)^n \det(\boldsymbol{\Sigma}) \big)$$

Building off the idea of entropy, one can begin to make statements about the dependence of two random variables by considering the conditional entropy of one given another.

**Definition 2.4.4.** ***Conditional Entropy and Differential Entropy*** [16]: *Let $X$ and $Y$ be random vectors with joint probability mass function (joint density) $p(x, y)$ and conditional $p(x|y)$. Then the conditional (differential) entropy of $X$ given $Y$ (when it exists) is defined*

$$H(X|Y) = -E_{p(x,y)} \big( \log_2 p(x|y) \big) \tag{2.4.3}$$

$H(X|Y)$ can be thought of as measuring the amount of disorder in $X$ that is not explained by $Y$. From the definitions it is clear that $H(X|Y) = H(X)$ if $X$ and $Y$ are independent. It is also easily shown that if $X = g(Y)$, where $g$ is a deterministic function, then $H(X|Y) = 0$. Because the logarithm turns multiplication to addition, Baye's rule has a nice analog with entropies

**Lemma 2.4.5.** ***Bayes' Rule for Entropies*** [16]: *If $X$ and $Y$ are random vectors then*

$$H(X, Y) = H(Y|X) + H(X)$$

**Corollary 2.4.6.** *If $X$ has mutually independent elements $\{X_i\}_{i=1}^n$ then*

$$H(X) = \sum_{i=1}^{n} H(X_i)$$

Finally, we mention a lemma that underlies much of the coming theory and applies to both regular and conditional entropies

**Lemma 2.4.7.** ***Differential Entropy Under Invertible Transformations*** [16]: *Let $X$ and $Y$ be random vectors satisfying $Y = g(X)$, where $g$ is an invertible transformation, and let $J$ be the Jacobian of $g$. Then*

$$H(Y) = H(X) + \log_2 |\det(J)|$$

*This result holds if the entropy is conditioned on another random variable, $Z$.*

**Proof:** We will prove the above result for the conditional entropy case as it generalizes the regular entropy. Let $X$ and $Y$ take values in $\Omega_y$, $Z$ be a random variable taking values on $\Omega_z$, let $p_{x|z}$ and $p_{y|z}$ be the conditional densities of $X$ and $Y$ w.r.t $Z$ respectively, let $p_{y,z}$ be the joint density of $Y$ and $Z$ and $p(z)$ be the density of $Z$. Assuming all quantities are well defined we proceed as follows

$$H(X|Z) = -\int_{\Omega_z}\int_{\Omega_y} \log_2\big(p_{x|z}(x)\big)p_{x|z}(x)p(z)dxdz$$

$$= -\int_{\Omega_z}\int_{\Omega_y} \log_2\Big(p_{y|z}\big(g(x)\big)|\det(J)|\Big)p_{y|z}\big(g(x)\big)|\det(J)|dxp(z)dz$$

$$= -\int_{\Omega_z}\int_{\Omega_y} \log_2\big(p_{y|z}(y)|\det(J)|\big)p_{y|z}(y)dyp(z)dz$$

$$= -\int_{\Omega_z}\int_{\Omega_y} \log_2\big(p_{y|z}(y)\big)p_{y,z}(y,z)dydz - \log_2|\det(J)|$$

$$= H(Y|Z) - \log_2|\det(J)|$$

∎

### Mutual Information

One frequently wants to know how much variability in one random variable is explained by the other, rather than how much variability in one is independent of the other, as is given by conditional entropy. This desire leads logically to the definition of mutual information

**Definition 2.4.8.** ***Mutual Information:*** *Let $X$ and $Y$ be random vectors. Then the mutual information between $X$ and $Y$ is defined*

$$I(X,Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

Using Bayes' rule for entropies shows that mutual information is symmetric. Unlike correlation, mutual information describes the dependence between random variables and is thus a very powerful tool. Mutual information is a special case of the Kullback Leibler Divergence (KLD), which is a measure of distance (although not quite a metric because it fails to satisfy symmetry) between probability distributions. We will define KLD because it provides extra insight into the idea of mutual information as a distance between probability distributions.

**Definition 2.4.9.** ***Kullback Leibler Divergence:*** *Let $X$ and $Y$ be discrete (continuous) random vectors with probability mass (density) functions $p_x(s)$ and $p_y(s)$. The KLD of the distributions of $X$ and $Y$ is*

$$D(p_x||p_y) = E_{p_x(s)}\left[\log_2\left(\frac{p_x(s)}{p_y(s)}\right)\right]$$

Substituting the definitions of entropy and conditional entropy into the definition of mutual information (2.4.8) shows that mutual information is simply $D\big(p(x,y)||p(x)p(y)\big)$, the divergence between the joint distribution of $X$ and $Y$ and what their joint distribution would be if they were independent. We will finish this section with several properties of mutual information and one regarding conditional entropy.

**Theorem 2.4.10.** ***Non-Negativity of Mutual Information*** [16]: *Let $X$ and $Y$ be random vectors with probability density (mass) functions given by $p_x$ and $p_y$ respectively, and joint distribution $p_{x,y}$. Then*

$$I(X;Y) \geq 0$$

**Proof:**

$$-I(X;Y) = E_{p_{x,y}}\left[\log_2\left(\frac{p_x p_y}{p_{x,y}}\right)\right] \leq \log_2\left(E_{p_{x,y}}\left[\frac{p_x p_y}{p_{x,y}}\right]\right)$$

$$= \log_2 \left( \int_{\Omega_x, \Omega_y} \frac{p_x p_y}{p_{x,y}} p_{x,y} dx dy \right) = \log_2(1) = 0$$

where we have used the definition of mutual information and Jensen's inequality. We see that $I(X;Y) \geq 0$ and we further note by Jensen's inequality that equality only occurs when $p_x p_y = p_{x,y}$, that is, when $X$ and $Y$ are independent. ∎

**Corollary 2.4.11.** ***Entropy Bounds Conditional Entropy*** [16]: *Let $X$ and $Y$ be random vectors. Then*

$$H(Y) \geq H(Y|X)$$

**Theorem 2.4.12.** ***Invariance to Invertible Transformations*** [42]: *Let $X$ and $Y$ be random vectors with values in $\mathbb{R}^n$, probability density (mass) functions given by $p_x$ and $p_y$, respectively, and joint distribution $p_{x,y}$. Let $g$ be an invertible transformation. Then*

$$I(X;Y) = I\big(g(X); g(Y)\big)$$

**Proof:** From the definition of mutual information

$$I\big(g(X); g(Y)\big) = H\big(g(X)\big) + H\big(g(Y)\big) - H\big(g(X), g(Y)\big)$$

$$= H(X) + H(Y) + 2\log_2\big(|\det(J)|\big) - H(X,Y) - \log_2\left(\left|\det\left(\begin{bmatrix} J & 0 \\ 0 & J \end{bmatrix}\right)\right|\right)$$

$$= I(X;Y) + 2\log_2\big(|\det(J)|\big) - \log_2\big(|\det(J)|^2\big) = I(X;Y)$$

where we have used lemma 2.4.7 for the second equality and, for the second last equality, the fact that the determinant of a block diagonal matrix is the product of the blocks.

∎

The following three lemmas are included without citation because, although they follow very simply from some of the above results, we could not find explicit statements of them anywhere. We expect that they are difficult to find in the literature because of their triviality.

**Lemma 2.4.13.** ***Mutual Information for Independent Elements I:*** *Let* $X = [X_1...X_n]^T$ *and* $Y = [Y_1...Y_n]^T$ *be random vectors s.t. the elements of* $X$ *are mutually independent. Then*

$$I(X;Y) \geq \sum_{i=1}^{n} I(X_i;Y_i)$$

**Proof:**

$$I(X;Y) = H(X) - H(X|Y) = \sum_{i=1}^{n} H(X_i|X_1,...,X_{i-1}) - H(X_i|X_1,...,X_{i-1},Y)$$

$$= \sum_{i=1}^{n} H(X_i) - H(X_i|X_1,...,X_{i-1},Y) \geq \sum_{i=1}^{n} H(X_i) - H(X_i|Y_i) = \sum_{i=1}^{n} I(X_i;Y_i)$$

where the second equality follows from *Bayes' Rule for Entropies* (corollary 2.4.5), the third equality from independence of $X_i$'s and corollary 2.4.6, and the single inequality from *Entropy Bounds Conditional Entropy* (lemma 2.4.11). ∎

**Lemma 2.4.14.** ***Mutual Information for Independent Elements II:*** *Let* $X = [X_1...X_n]^T$ *and* $Y = [Y_1...Y_n]^T$ *be random vectors and let* $\{Z_i = (X_i, Y_i) \mid i \in 1,...,n\}$. *Further assume the elements of* $X$ *are mutually independent, the elements of* $Y$ *are mutually independent and the elements of the set* $\{Z_i\}_{i=1}^n$ *are as well. Then*

$$I(X;Y) = \sum_{i=1}^{n} I(X_i;Y_i)$$

**Proof:** Follows from corollary 2.4.6. ∎

**Lemma 2.4.15.** ***Mutual Information for Independent Elements III:*** *Let* $X = [X_1...X_n]^T$ *and* $Y = [Y_1...Y_n]^T$ *be random vectors and let* $Y_X$ *be the random vector of elements of* $Y$ *given* $X$ *is known. Assume that the elements of* $Y$ *are mutually independent and the elements of* $Y_X$ *are as well. Then*

$$I(X;Y) = \sum_{i=1}^{n} I(Y_i;X)$$

**Proof:** Follows from corollary 2.4.6. ∎

**Theorem 2.4.16.** ***Data Processing Inequality*** [16]: *Let $X$, $Y$ and $Z$ be random vectors such that $X \to Y \to Z$ defines a Markov chain. Then*

$$I(X;Y) \geq I(X;Z)$$

**Proof:**

$$X \to Y \to Z \implies p(x|y,z) = p(x|y)$$

Using the definition of entropy this implies $H(X|Y,Z) = H(X|Y)$, which gives the following

$$I(X;Y,Z) = H(X) - H(X|Y,Z) = H(X) - H(X|Y) = I(X;Y)$$

It remains to prove that $I(X;Y,Z) \geq I(X;Z)$. Note that $H(X|Z) - H(X|Y,Z)$ is, by the definition of mutual information, the information between $X$ and $Y$ given $Z$, which we denote $I(X;Y|Z)$. Using the definition of mutual information and Bayes' rule for entropies (lemma 2.4.5) we have

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \implies I(X;Y,Z) \geq I(X;Z)$$

by non-negativity of mutual information (Th. 2.4.10). ∎

**Theorem 2.4.17.** ***Gaussian Distribution Maximizes Entropy on*** $\mathbb{R}^n$ [16]: *Let $\mathcal{X}$ be the set of all random vectors on $\mathbb{R}^n$ with covariance function $\Sigma$. Then* $\max\{H(x) \mid x \in \mathcal{X}\}$ *is a normal distribution.*

### 2.4.2 Entropy and Information Rates

A natural extension of entropy and information to stochastic processes are the entropy and information rates, the respective quantity per unit time of the process. However, if the statistics of a process change over time then there is no reason why the entropy or information rate should not change also. Accordingly, these objects only make sense for describing stationary processes. We will begin by defining the entropy and mutual information rates and then discuss under what conditions the two quantities exist.

**Definition 2.4.18.** ***Entropy and Information Rates:*** *Let $\{X_t\}_{t\in\mathbb{N}}$ and $\{Y_t\}_{t\in\mathbb{N}}$, be stationary stochastic processes. Then the entropy rate of $X$ is*

$$\mathbb{H}(X) = \lim_{T\to\infty} \frac{H(X_1,...,X_T)}{T}$$

*and the mutual information rate between $X$ and $Y$ is*

$$\mathbb{I}(X;Y) = \lim_{T\to\infty} \frac{I(X_1,...,X_T;Y_1,...,Y_T)}{T}$$

The following result, which we will state without proof, allows one to show existence of entropy rate

**Theorem 2.4.19.** ***Barron's Ergodic Theorem*** [5]: *If $X$ is a stationary, ergodic stochastic process with $T^{th}$ finite dimensional distribution given by $f(X_1, ..., X_T)$ then*

$$\lim_{T\to\infty} -\frac{\log_2\big(f(X_1, ..., X_T)\big)}{T} = \lim_{T\to\infty} H(X_T|X_1, ..., X_{T-1}) < \infty \quad a.s.$$

The following two corollaries give well known results; a complete discussion of the convergence of entropy and information rates can be found in the text Entropy and Information, by Gray [33].

**Corollary 2.4.20.** *If $X$ is a stationary, ergodic stochastic process then its entropy rate exists and is equal to $\lim_{T\to\infty} H(X_T|X_0, ..., X_{T-1})$.*

**Proof:** Let $(\Omega, \mathcal{F}, P)$ be the probability space on which $X$ is defined and let $f$ be the finite dimensional distribution as in Barron's theorem

$$\mathbb{H}(X) = -\lim_{T\to\infty} \int_\Omega \frac{1}{T} \log_2\big(f(X_0, ..., X_T)\big) dP(\omega)$$

Using the convergence property given by Barron's Theorem we can bound the integrand by some constant $M$. Because $P$ is a probability measure it integrates to 1 so $M$ is integrable w.r.t. to $P$ and we can apply dominated convergence theorem (Th. 2.1.10) to bring the limit inside the integral.

$$\mathbb{H}(X) = -\int_\Omega \lim_{T\to\infty} \frac{1}{T} \log_2\big(f(X_0, ..., X_T)\big) dP(\omega)$$

$$= \int_\Omega \lim_{T\to\infty} H(X_T|X_0, ..., X_{T-1}) dP(\omega)$$

The integrand is, by Barron's theorem, less than infinity almost surely and we are integrating with respect to a probability measure so the integral is finite. ∎

One can similarly define a mutual information rate, a result first due to Dobrushin [33]

**Corollary 2.4.21.** *If $X$ and $Y$ are jointly stationary, ergodic stochastic processes then their mutual information rate exists and is given as follows*

$$\mathbb{I}(X;Y) = \lim_{T\to\infty} \big[H(Y_T|Y_0, ..., Y_{T-1}) - H(Y_T|Y_0, ..., Y_{T-1}, X_{0:T})\big]$$

*where $X_{0:T}$ denotes $[X_0...X_T]$.*

**Proof:** Let $Z$ be the stochastic process such that $Z_t = (X_t, Y_t)$. If $X$ and $Y$ are jointly stationary then $Z$ is also because its finite dimensional distributions are shift invariant. Using the definition of mutual information we have

$$\mathbb{I}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(Z)$$

By the previous corollary each of the three terms exist so $\mathbb{I}$ does also. By applying Bayes' rule for entropies (lemma 2.4.5) we arrive at the result. ∎

The following theorem, due to Kolmogorov [16], gives the entropy rate for a stationary Gaussian process.

**Theorem 2.4.22.** ***Entropy Rate for Stationary Discrete Gaussian Process:*** *Let $X$ be a univariate, stationary, discrete Gaussian process with power spectrum $\mathcal{P}_{XX}(f)$.*

$$\mathbb{H}(X) = \int_0^{\frac{1}{2}} \log_2\big(\mathcal{P}_{XX}(f)\big)df + \frac{1}{2}\log_2(2\pi e)$$

The following theorem attempts to provide theoretical justification for the correlation theory result of [21] and will also be used as a key part of the primary mathematical contribution of this thesis. However, it appears to be incorrect as written because it is missing a factor of two in the additive constant term (which is necessary to coincide with the above result regarding Gaussian processes).

**Lemma 2.4.23.** ***Entropy Rate as an Integral:*** *Let $X$ be a discrete, stationary stochastic process, let $\hat{X}^{(T)} = [\hat{X}_{f_0}^{(T)}...\hat{X}_{f_T}^{(T)}]$ be the DFT of a finite sample of $X$, where the frequencies are defined as in appendix C.2, and let $\hat{X}$ be the limiting distribution, given by Brillinger's theorem, of the DFT of $X$. If the convergence of the Fourier transform is well enough behaved that $\lim_{T\to\infty} \frac{H(\hat{X}^{(T)}) - H(\hat{X})}{T} \to 0$ then*

$$\mathbb{H}(X) = \int_0^{\frac{1}{2}} \log_2\big(\mathcal{P}_{XX}(f)\big)df + \frac{1}{2}\log_2(\pi e)$$

**Proof:** The DFT is an orthogonal transformation so its Jacobian is equal to 1 and the density in the time domain is equal to the density in the frequency domain. Let $2N_T + \alpha = T$ (subscript denotes dependence on $T$), with $\alpha = \{0, 1\}$ depending on whether the length of the series is odd or even. The sample spacing in the frequency domain is $\Delta f_T = \frac{1}{T}$ (the subscript $T$ again denotes dependence on $T$). Using these two points yields the following

$$\mathbb{H}(X) = \lim_{T\to\infty} \frac{H(X_1, X_2, ..., X_{T-1}, X_T)}{T} = \lim_{T\to\infty} H(\hat{X}_{f_0}^{(T)}, ..., \hat{X}_{f_{2N_T+\alpha}}^{(T)})\Delta f_T$$

Using the assumption about the convergence of entropies, the independence structure for the DFT given by Brillinger's proof (Th. 2.3.5) and the order of DFT frequencies described in appendix C.2 gives

$$= \lim_{T\to\infty} \left[H(X_{f_0}) + (1-\alpha)H(\hat{X}_{f_{N_T}}) + \sum_{n=1}^{N_T-(1-\alpha)} H(\hat{X}_{f_n}, \hat{X}_{f_{2N_T+\alpha-n}})\right]\Delta f_T$$

$$= \lim_{N\to\infty} \sum_{n=0}^{N} H(\hat{X}_{f_n})\Delta f_N = \int_0^{\frac{1}{2}} H(\hat{X}_f)df$$

where the second last equality follows from two points: (1) we have switched the limit from $T$ to $N$ and simplified things by defining $N_T = N$ and $f_N = f_T$; (2) because the negative frequencies in the DFT are simply the complex conjugates of the positive frequencies they are statistically equivalent, that is $H(\hat{X}_{f_n}, \hat{X}_{f_{-n}}) = H(\hat{X}_{f_n})$ using Bayes' rule for entropies (Th. 2.4.5). The last equality uses the assumption that $H(\hat{X}_{f_n})$ is a continuous function of $f$ on all but a measure zero set of $\mathbb{R}$, which will be shown to be justified below, allowing the right hand side to be written as an integral using the definition of the Riemann integral. The entropy within the integral is given by Brillinger's theorem

$$= \int_0^{\frac{1}{2}} \frac{1}{2}\log_2\left((2\pi e)^2 \det(\boldsymbol{X})\right)df$$

$$\boldsymbol{X} = \frac{1}{2}\begin{bmatrix} \mathcal{P}_{XX}(f) & 0 \\ 0 & \mathcal{P}_{XX}(f) \end{bmatrix}$$

The result follows trivially by calculating the determinant and simplifying. It can be shown that $\mathcal{P}_{XX}(f)$ is a continuous function of $f$ [11]. Thus $H(\hat{X}_{f_n})$ is indeed a continuous function of $f$ on $[0, \frac{1}{2}]$, providing the justification alluded to above and completing the proof. ∎

## 2.5 Estimating Information Rates

In this final section of background two methods of calculating information rate will be discussed. Given the definitions of the previous section one could calculate information rate between stochastic processes $X$ and $Y$ directly without any further

insights. This approach amounts to estimating the joint distribution of $X^{(T)}$ and $Y^{(T)}$, where $X^{(T)} = [X_0...X_T]$ for stochastic process $X$, and $Y^{(T)}$ is defined equivalently, and $T$ must be large so as to approximate the limit in the definition of information rate. Estimation of this joint distribution will require excessive amounts of data even in the case where $X$ and $Y$ are univariate stochastic processes. An alternative estimation method is thus required to avoid running prohibitively long neural network model simulations for the collection of enough data to directly estimate an information rate.

The search for alternative methods of information rate estimation has a long history in theoretical neuroscience because information rate is an ideal tool for relating stimulus signals to neural responses. We will make use of methods developed out of this tradition: the linear lower bound [70], popularized by Bialek [60], and the Correlation Theory method [21]. The primary mathematical contributions of this thesis are to (1) realize that the latter approach can be applied not solely to spike-trains but to other kinds of stationary signals; (2) provide a more rigorous proof of Correlation Theory than is given in the original paper; (3) suggest that the lower bound will be equal to the full information rate for jointly stationary processes that satisfy the convergence of entropies required in lemma 2.4.23.

### 2.5.1 Lower Bound

The lower bound method, developed by Stein [70], is a cornerstone methodology for theoretical neuroscience and an important result for signal processing more generally. It demonstrates that, when the input signal is Gaussian, a lower bound on the mutual information between the input and the output can be obtained from the coherence of the two signals. The lower bound becomes equal to the full information when the input is related to the output by a linear transfer function with additive noise. The lower bound was also shown to be the linear approximation of the correlation theory [21] method for calculating mutual information. Calculating the lower bound is efficient because it is an estimation of spectral densities, rather than of a full distribution, and is rapidly calculated using standard signal processing algorithms [78]. We provide the lower bound result and, for its importance to this thesis, a review of the result's proof below.

**Theorem 2.5.1.** ***Information Rate Lower Bound*** [28]: *Let $X$ be a univariate Gaussian stochastic process, $Y$ be a univariate stochastic process and $\Phi_{XY}$ denote the coherence of the two processes at frequency $f$. Then $\mathbb{I}(X;Y) \geq l(X;Y)$, where $l$ is defined*

$$l(X;Y) = -\int_0^{\frac{1}{2}} \log_2\big(1 - \Phi_{XY}(f)\big)df$$

*This linear bound will be equal to the information when $Y = T * X + \xi$ where $T$ is a linear filter and $\xi$ is noise.*

**Proof:** First, note that because mutual information is invariant to invertible transformations (Th. 2.4.12) it is invariant to changes of basis. For this reason we can calculate mutual information using the frequency domain representations of the stochastic processes. For the rest of the proof when we refer to a given stochastic process we will be considering its frequency domain representation. Let $\hat{X}$ denote frequency representation of $X$. Let $X_{\text{est}} = T * Y$ be the optimal mean squares linear approximation of $X$ from $Y$ and let $N$ be the noise process defined by $N = X - X_{\text{est}}$. By the data processing inequality (Th. 2.4.16)

$$I(\hat{X};\hat{Y}) \geq I(\hat{X};\hat{X}_{\text{est}}) = H(\hat{X}) - H(\hat{X}|\hat{X}_{\text{est}}) = H(\hat{X}) - H(\hat{N} + \hat{X}_{\text{est}}|\hat{X}_{\text{est}})$$

by the definition of mutual information and the definition of $N$. Note that $H(\hat{N} + \hat{X}_{\text{est}}|\hat{X}_{\text{est}}) = H(\hat{N}|\hat{X}_{\text{est}}) \leq H(\hat{N})$ by lemma 2.4.7 (Differential Entropy under Invertible transformations) and by lemma 2.4.5 (Bayes' Rule for Entropies). Let $N_{\mathcal{N}}$ be the Gaussian process with frequency domain covariance function equal to that of $N$. By theorem 2.4.17 (Normal Distribution Maximizes Entropy...) $H(\hat{N}) \leq H(\hat{N}_{\mathcal{N}})$. Making use of the expression for the entropy rate of Gaussian processes (Th. we get

$$I(\hat{X};\hat{Y}) \geq H(\hat{X}) - H(\hat{N}_{\mathcal{N}}) = -\int_0^{\frac{1}{2}} \log_2 \left( \frac{\mathcal{P}_{NN}(f)}{\mathcal{P}_{XX}(f)} \right) df$$

where $\mathcal{P}_{XX}$ and $\mathcal{P}_{NN}$ are the power spectra of the processes $X$ and $N$ respectively. We now focus on the noise to signal power spectrum ratio inside the logarithm in the last equation. It can be shown [58, 28] that, if $T$ is the linear filter that minimizes noise in the mean square sense, $\hat{T}(f) = \frac{\mathcal{P}_{XY}(-f)}{\mathcal{P}_{YY}(f)}$. Using this observation, the definition of $N$ and the fact that the cross spectrum for real processes is symmetric completes the proof:

$$\frac{\mathcal{P}_{NN}(f)}{\mathcal{P}_{XX}(f)} = \frac{\mathcal{P}_{XX}(f) - \hat{T}(f)\hat{T}(f)^*\mathcal{P}_{YY}(f)}{\mathcal{P}_{XX}(f)} = 1 - \frac{\mathcal{P}_{XY}(f)^*\mathcal{P}_{XY}(f)}{\mathcal{P}_{XX}(f)\mathcal{P}_{YY}(f)} = 1 - \Phi_{XY}(f)$$

∎

### 2.5.2 Correlation Theory

The correlation theory method [21] was inspired by the asymptotic normality of frequency components and the empirical observation that the output spike-train of a model neuron given a fixed stimulus sample path also appears asymptotically normal. Based on these two phenomena an expression for the full mutual information between the input to a single neuron and the output spike train was derived. We noticed that the correlation theory result could be readily applied to calculate the

mutual information between any two stationary signals, so long as the asymptotic normality of the Fourier components of one signal given a sample path from the other signal could be verified as in the original paper. This more general statement of correlation theory is provided below. The formula provided here is essentially the same as in the original work; however, the assumptions required to extend the result to stochastic processes more generally and the proof of the result directly via the definition of the circularly symmetric complex normal are contributions of this thesis.

**Theorem 2.5.2.** ***Correlation Theory*** [21]: *Let $X$ and $Y$ be stationary stochastic processes and let the stochastic process $Y_x$, denote the process $Y$ given a fixed sample path of $X$. If $Y_x$ has DFT at frequency $f$ given by $\hat{Y}_{f,\hat{x}} \sim \mathcal{N}_{\mathbb{C}}(\mu, \Sigma_{\hat{x}})$ such that DFTs at different frequencies are independent and if the entropy $H(Y_x)$ is a bounded, continuous function of $f$ on $[0, \frac{1}{2}]$ then the information rate between $X$ and $Y$ is*

$$\mathbb{I}(X;Y) = \int_0^{\frac{1}{2}} E_x\left[\log_2\left(\frac{\sigma^2(f)}{\sigma^2_{\hat{x}}(f)}\right)\right] df$$

*where the expected value is over sample paths of $X$, $\sigma^2(f)$ and $\sigma^2_{\hat{x}}(f)$ are the variances of the real and imaginary parts of $\hat{Y}$ and variance of the real and imaginary parts of $\hat{Y}_{\hat{x}}$, respectively.*

**Proof:** Because mutual information is unaffected by change of basis we can switch to the frequency domain (Th. 2.4.12). Because both entropy rate and conditional entropy rate are finite, by assumptions, we write the information rate as a difference. By the assumptions about $Y_x$ we can use the approach used in equalities 3, 4 and 5 in the proof of lemma 2.4.23 to write both entropies as integrals

$$\mathbb{I}(X;Y) = \mathbb{H}(\hat{Y}) - \mathbb{H}(\hat{Y}|\hat{X}) = \int_0^{\frac{1}{2}} H(\hat{Y}_f) - H(\hat{Y}_f|\hat{X}) df \tag{2.5.1}$$

Now focus on the integrand. This is the mutual information between $X$ and each frequency $f$ of $Y$: $I(\hat{X};\hat{Y}_f)$.

$$H(\hat{Y}_f) - H(\hat{Y}_f|\hat{X}) = H(\hat{Y}_f) - E_{\hat{x}}\big[H(\hat{Y}_{f,\hat{x}})\big] = E_{\hat{x}}\big[H(\hat{Y}_f) - H(\hat{Y}_{f,\hat{x}})\big] \tag{2.5.2}$$

We now evaluate the expression inside the expected value in (2.5.2). Again by the assumptions about $Y_x$ we can proceed as in the final equality in the proof of 2.4.23 to write both entropies using the definition for multivariate Gaussian entropy. Note however that Brillinger's result about the variance of the DFT being equal to the power spectrum need not apply, so we will write the resulting expression as a fraction of variances instead of power spectra.

$$I(\hat{X};\hat{Y}_f) = E_x\left[\log_2\left(\frac{\sigma^2(f)}{\sigma^2_{\hat{x}}(f)}\right)\right] \tag{2.5.3}$$

Inserting this expression as the integrand in equation (2.5.1) gives the result.

∎

The statement of Correlation Theory in this thesis is different than the original in that the equation is written as a fraction of variances instead of expanding into a Fourier transform of autocorrelation functions, as in the original paper. This simpler statement of Correlation Theory was adopted for two reasons. First, we are uncertain whether the full Correlation Theory result can be proven rigorously; second, we found that estimation of information rate using the original equation was unwieldy from a computational perspective.

The proof of Correlation Theory here also diverges slightly from the original. The original proof used the relationship between the univariate, circularly symmetric complex normal random variable and the Rayleigh distribution. The proof given here uses the circularly symmetric complex normal directly and, hence, arrives at the result more quickly.

### 2.5.3 Lower Bound as Full Information Rate for Certain Processes

If the two stochastic process arguments for information rate are jointly stationary and, critically, the extra assumption that the entropy rate convergence requirement used in lemma 2.4.23 holds, it can be shown that the lower bound from theorem 2.5.1 is equal to the full information rate. This result is, to the best of my knowledge, novel.

**Lemma 2.5.3.** ***Lower Bound as Full Information Rate:*** *Let $X$ and $Y$ be jointly stationary, finite memory stochastic processes satisfying the assumptions of lemma 2.4.23. Then the full mutual information rate between $X$ and $Y$ is given by the lower bound.*

**Proof:** Let $Z = [X\,Y]^T$ be the multivariate stochastic process formed by concatenating $X$ and $Y$. Because of the joint stationarity of the two original processes $Z$ is also stationary. We can thus apply lemma 2.4.23 to all three processes:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(Z) = \int_0^{\frac{1}{2}} H(\hat{X}_f) + H(\hat{Y}_f) - H(\hat{Z}_f)df \tag{2.5.4}$$

where we have also used the linearity of the integral and defined $\hat{X}_f$ to be the limiting distribution of the DFT of $X$ (as in lemma 2.4.23) at frequency $f$. By Brillinger's theorem (Th. 2.3.5) each of $\hat{X}_f$, $\hat{Y}_f$ and $\hat{Z}_f$ are complex normal random variables. Complex random variables can be represented as real multivariate random variables with double the original dimension. Using these observations, the definition of the complex normal random variable (section 2.3.3) and the entropy for multivariate normals (lemma 2.4.3) allows rewriting the RHS of equation 2.5.4 as

$$= \int_0^{\frac{1}{2}} \frac{1}{2}\log_2\left((2\pi e)^2 \det(\boldsymbol{X})\right) + \frac{1}{2}\log_2\left((2\pi e)^2 \det(\boldsymbol{Y})\right) - \frac{1}{2}\log_2\left((2\pi e)^4 \det(\boldsymbol{Z})\right) df$$

$$\boldsymbol{X} = \frac{1}{2}\begin{bmatrix} \mathcal{P}_{XX}(f) & 0 \\ 0 & \mathcal{P}_{XX}(f) \end{bmatrix}$$

$$\boldsymbol{Y} = \frac{1}{2}\begin{bmatrix} \mathcal{P}_{YY}(f) & 0 \\ 0 & \mathcal{P}_{YY}(f) \end{bmatrix} \tag{2.5.5}$$

$$\boldsymbol{Z} = \frac{1}{2}\begin{bmatrix} \mathcal{P}_{XX}(f) & \mathrm{Re}\big(\mathcal{P}_{XY}(f)\big) & 0 & -\mathrm{Im}\big(\mathcal{P}_{XY}(f)\big) \\ \mathrm{Re}\big(\mathcal{P}_{XY}(f)\big) & \mathcal{P}_{YY}(f) & \mathrm{Im}\big(\mathcal{P}_{XY}(f)\big) & 0 \\ 0 & \mathrm{Im}\big(\mathcal{P}_{XY}(f)\big) & \mathcal{P}_{XX}(f) & \mathrm{Re}\big(\mathcal{P}_{XY}(f)\big) \\ -\mathrm{Im}\big(\mathcal{P}_{XY}(f)\big) & 0 & \mathrm{Re}\big(\mathcal{P}_{XY}(f)\big) & \mathcal{P}_{YY}(f) \end{bmatrix}$$

evaluating these determinants and moving the exponent inside each logarithm to the outside gives

$$= \int_0^{\frac{1}{2}} \log_2\left(\pi e \mathcal{P}_{XX}(f)\right) + \log_2\left(\pi e \mathcal{P}_{YY}(f)\right) - \log_2\left((\pi e)^2(\mathcal{P}_{XX}(f)\mathcal{P}_{YY}(f) - |\mathcal{P}_{XY}|^2(f))\right) df$$

$$= \int_0^{\frac{1}{2}} \log_2\left(\frac{\mathcal{P}_{XX}(f)\mathcal{P}_{YY}(f)}{\mathcal{P}_{XX}(f)\mathcal{P}_{YY}(f) - |\mathcal{P}_{XY}(f)|^2)}\right) df$$

$$= -\int_0^{\frac{1}{2}} \log_2\left(1 - \Phi_{XY}(f)\right) df \tag{2.5.6}$$

where $\Phi_{XY}(f)$ is the coherence function.

∎

The significance of the above result is that it extends the use of the lower bound, which, in comparison to correlation theory, does not require numerical validation of asymptotic normality for non-stationary time series and is substantially more data efficient. However, like correlation theory, it relies on the entropy rate convergence assumed in lemma 2.4.23, which remains to be proven.

# Chapter 3

# Results

In the following we will discuss the information in three channels: (i) the burst channel, composed of burst input and decoded burst signal output; (ii) the event channel, comprising event input and decoded event signal output; (iii) the burst multiplexing channel. The multiplexing channel could be defined in two ways: either as the information between the bivariate input process, composed of burst and event inputs, and bivariate output process, or as the sum of the information in the burst and event channels. The latter option was chosen for the following reasons. From the independence of the input processes used and lemma 2.4.13 the sum of the channels will not accidentally count information twice as it is bounded by the multivariate information rate. In the likely case that the sum of the channels is less than the information in the bivariate process the lost information will be information in the decoded burst signal about the input event signal and in the decoded event signal about the input burst signal. Because the goal of burst multiplexing is to separate these two signals as best as possible this extra information between the bivariate processes is due to errors in the burst multiplexing code and is thus noise. For this reason the sum of information in the burst and event channels represents a better characterization of the multiplexing channel than the bivariate method.

The primary result of this thesis is that information transmission in the burst, event and burst multiplexing channels is robust to changes in ISI distribution shape. To get to this result $\text{SRM}_0^2$ was simulated using parameters chosen to best reflect the biology of the modeled system (see Appendix A.1 for full parameter details and values). However, several decoding network parameters required preliminary investigation to set. These were the decoding function parameters, the power spectra for the input signals, and the lag between burst and event signal used in calculating the decoded burst fraction. We discuss our method of selecting these hyper-parameters before exploring the primary results.

## 3.1 Calibration

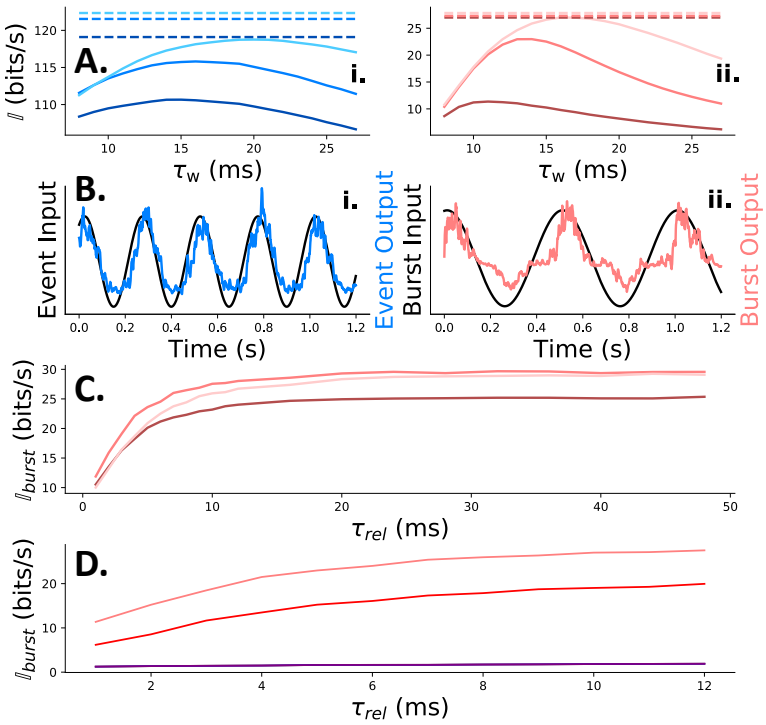


Figure 3.1: **Hyper-Parameter Adjustments:** **A.** Search for optimal weight function parameter in imperfectly classified spike regime: the information rate between input and output signals is plotted as a function of weight function parameter for three ISI distribution shapes (darker to bright colour denotes increasing relative refractory time constant). The dashed lines are the mutual information for perfect spike classification for each ISI distribution shape **i.** Event channel **ii.** Burst channel **B.** Signal decoding with perfectly classified spikes: the input (black line) is decoded by the synaptic potential of the post-synaptic cell (coloured line). The event input is chosen to have higher frequencies than the burst input. **i.** Event input (black) is decoded relatively well by event rate decoding synaptic potential (blue). **ii.** Burst input (black) is decoded, with noticeable noise from the event channel, by burst fraction decoding synaptic potential (orange). **C.** Lower bound on information rate as a function of relative refractory period time constant (ISI distribution shape) for three different lags between burst and event signals in burst fraction quotient. 0 ms lag is light orange, 13 ms is dark and 6 is intermediately coloured. **D.** Demonstration that burst fraction outperforms other parameters, at least for linear decoding, in estimating burst input: lower bound on information rate is plotted as a function of ISI distribution shape. The purple line represents information between burst input and population activity, population event rate and the inverse population event rate; red line represents burst rate and orange line burst fraction.

First, it was important to choose weight function parameters to maximize the information rate for every ISI distribution shape to ensure a fair comparison across models with different distributions. Weight functions depend solely on the threshold parameter, $\tau_w \in \mathbb{R}$, which was optimized using a grid search. Information rate in burst and event channels is shown as a function of threshold parameter for three ISI distribution shapes in figure 3.1.A. There was some variability in the optimal thresholds for different ISI distributions, suggesting that short term plasticity rules used by a post-synaptic cell should change as a function of pre-synaptic spiking statistics if the cells are to maximize information transmission.

The second pair of hyper-parameters to set are those associated with the input signals. The burst signal will have greater difficulty encoding high frequency signals in finite encoding populations because the randomness in the intra-burst interval introduces extra uncertainty about rapidly occurring fluctuations in the input signal. For example, if the input signal is composed of two impulses the variability of the intra-burst ISI could cause the response to the first impulse to occur more rapidly, via shorter intra-burst ISIs, and the response to the second to occur later, via longer intra-burst ISIs. If the impulses occur close enough together this noise could make the impulses indistinguishable. For lower frequency impulses it is much less likely for this ambiguity to occur. For this reason, the burst input was chosen to have more power at lower frequencies, and less at higher frequencies, compared to the event input signal (see Appendix A.1). This difference in input frequency is illustrated in figure 3.1.B with deterministic sine waves rather than O-U processes, as used in the other results, to give intuition and to demonstrate the difference in decoding accuracy of the event and burst channels.

The intra-burst ISI also introduces another problem for decoding. To remove the effects of the event signal on the decoded burst signal one must normalize the number of bursts that occur at a given time by the number of events, i.e. calculate the burst fraction. However, the numerator of this fraction, the burst rate, is calculated from the second spike in the burst, the burst spike, while the denominator, event rate, is calculated by the first spike, the event spike. These two spikes are separated by an intra-burst ISI. This intra-burst ISI introduces a lag between the decoded event and burst rates that could hypothetically make decoding more difficult if not accounted for by shifting (introducing a lag in) the decoded event rate before calculating the burst fraction. Further motivation for exploring the effect of a lag in the burst fraction calculation comes from a hypothesis about how burst fraction could be computed in the brain [54]: that event rate could be decoded by one cell which could then divisively inhibit another cell whose synapses decode burst rate so that this second cell would encode burst fraction. This architecture would naturally introduce a lag in the event signal by having it pass through another cell before being used in the burst fraction. To explore the effect of lag on burst fraction estimation we calculated the lower bound on mutual information for three event signal lags (figure

3.1.C). These lags were chosen to be 0, 6 and 13 ms because the expected value of the intra-burst ISI is approximately 6 ms. As expected, the 6 ms lag was the best of the three tested and was thus used for all the primary results. A more precise optimization of lag was not carried out as it did not appear to effect results qualitatively over the tested range of values.

## 3.2 Primary Results

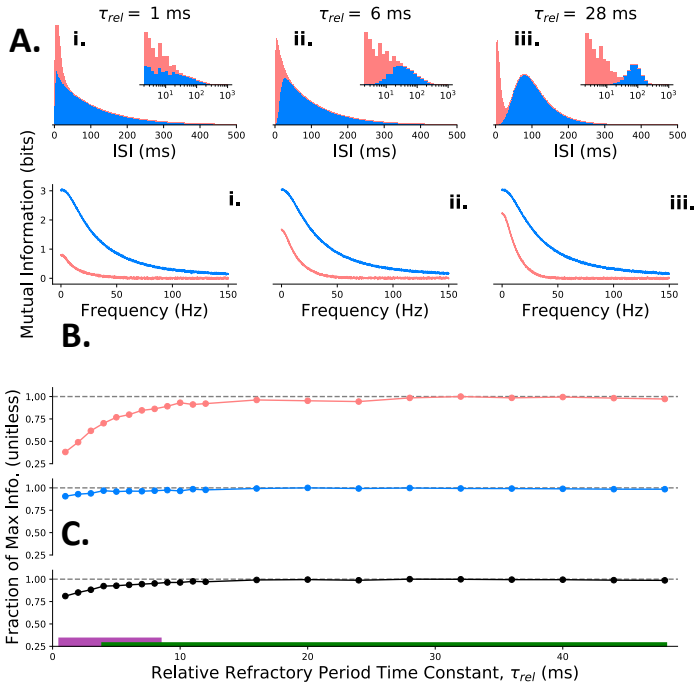


Figure 3.2: **Burst Multiplexing is Robust to ISI Shape:** Mutual information rate is minimally effected by ISI distribution shape. **A.** ISI distributions for three values of relative refractory period time constant: **i.** 1 ms, **ii.** 6 ms, **iii.** 28 ms **B.** Mutual information as a function of frequency for encoding populations with the ISI distributions in (A). **C.** Fraction of maximum information rate (max taken over all tested ISI distribution shapes) as a function of relative refractory period time constant (ISI distribution shape) in **i.** the burst channel (orange) **ii.** the event channel (blue) **iii.** the sum of channels. Dashed line indicates ratio of 1, green region (at bottom) denotes parameter range over which more than 70% of the max information is transmitted and purple region denotes parameter range for unimodal ISI distributions; note the overlap.

In contrast to the original hypothesis of this thesis, that the information transmitted under a burst multiplexing code would exhibit a clear maximum for some ISI distribution, we found that the burst multiplexing code is robust to changes in ISI distribution shape in the sense that many types of ISI distribution shapes,

ranging from bimodal to unimodal, are capable of effectively transmitting information. Indeed, all bimodal ISI distribution shapes and more than half of the unimodal ISI distributions performed at over 70% of the maximal information rate (see figure 3.2.C.i), the max being taken over all tested ISI distributions. From a minimal value at the smallest refractory period time constant, the information transmitted increased rapidly to a steady, maximum level in the bimodal ISI distribution regime (the bimodal regime beginning at $\tau_{rel} = 9$ms in figure 3.2.C). Information transmission was even less effected in the event channel, with the minimally transmitting ISI shape at more than 90% of the max value (see figure 3.2.C.ii).

As one would expect, given fewer burst spikes occur than event spikes, there is less information transmitted in the burst channel than in the event channel. For our study we set the mean burst probability to approximately 0.2 and, accordingly, observed an information rate in the burst channel that was roughly this fraction of the total burst multiplexing channel information: the ratio of maximal burst channel information to maximal burst multiplexing channel information (maximum being over ISI shapes) was approximately 0.18. Given this ratio the full burst multiplexing channel exhibited behaviour closer to the event channel in that the minimally performing ISI distribution shape still transmitted over 80% of the information rate of the maximally performing ISI distribution (see figure 3.2.C.iii.).

We had to verify numerically that the output time series given the input time series displayed asymptotic normality and independence properties (see theorems 2.3.5 and 2.5.2), for the figure 3.2 results, to validate the assumptions of Correlation Theory, which was the estimator used to produce the figure. This was done and the plots are shown in Appendix B.2. As a second check we recalculated mutual information using the lower bound (see theorem 2.5.1 and lemma 2.5.3) and found that they agreed with the Correlation Theory results.

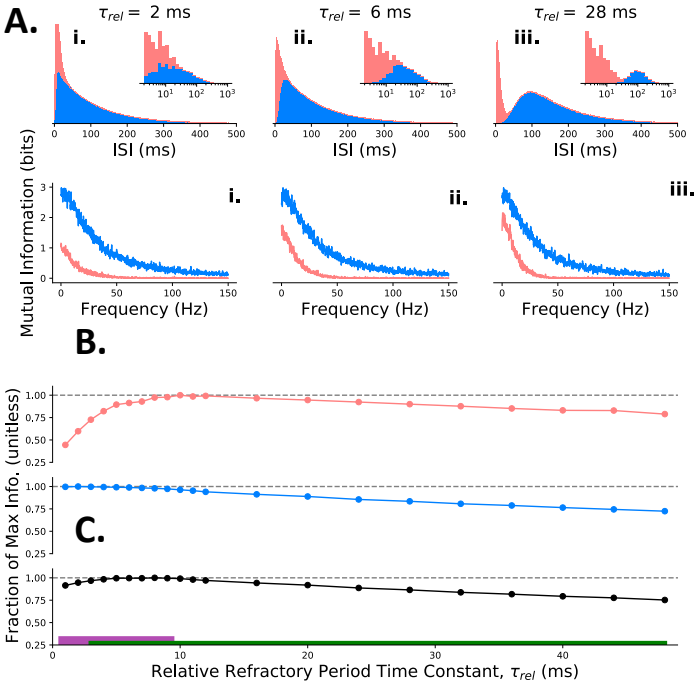


Figure 3.3: **Weakly Bimodal ISI Performs Best at Linear Decoding for Uncorrected Rate:** When firing rate is not corrected for information rate drops with greater separation between ISI distribution modes **A.** ISI distributions for three values of relative refractory period time constant: **i.** 2 ms, **ii.** 6 ms, **iii.** 28 ms **B.** Mutual information as a function of frequency for encoding populations with the ISI distributions in (A). **C.** Fraction of maximum information rate (max taken over all tested ISI distribution shapes) as a function of relative refractory period time constant (ISI distribution shape) in **i.** the burst channel (orange) **ii.** the event channel (blue) **iii.** the sum of channels. Dashed line indicates ratio of 1; green region (at bottom) denotes range over which 70% of the maximum information is transmitted; purple region denotes parameter range for unimodal ISI distributions. Information rate was estimated via lower bound.

Figure 3.2 was generated by matching spike counts for each ISI distribution shape. Unless a correction in the rate of event spikes was made, forcing a more bimodal ISI distribution would lead to event spikes occurring less frequently and thus lower power in the decoded signals. To fairly compare information transmitted for different ISI shapes it was necessary to increase the rate of event spikes as the

relative refractory period time constant, parameterizing the unimodal-bimodal ISI shape transition, was increased. The rate correction made for 3.2 is demonstrated in Appendix B.1, with the rates plotted as a function of relative refractory period to show that the correction was properly implemented. It remained unknown, however, how rapidly information would drop if rate was not corrected for, so the main experiment was repeated with uncorrected rates, the results of which are shown, utilizing the lower bound method, in 3.3.

In the rate corrected case the increase in relative refractory period time constant leads to a rapid decrease in information rate for all but the smallest values of the ISI shape parameterizing time constant. Because of this decrease there is an ISI distribution shape that maximizes information transmission over the tested ISI distribution parameters. This maximum occurs in the unimodal regime for the entire burst multiplexing channel and event channel but in the moderately bimodal regime for the burst channel.

# Chapter 4

# Discussion

This thesis will conclude with a discussion of the study assumptions and future directions in applied mathematics followed by an outline of implications and future directions in neuroscience.

### 4.0.1 Assumptions & Future Directions in Applied Math

For the lower bound on information rate to equal the full information for jointly stationary processes it is required that $\lim_{T\to\infty} \frac{H(\hat{X}^{(T)})-H(\hat{X})}{T} \to 0$, as assumed in lemma 2.4.23. Convergence of entropies is a non-trivial matter and is not implied by convergence in distribution [57], which is the form of convergence given by the Brillinger theorem. An important direction for future studies will be to investigate under what conditions this convergence in entropy rates occurs. A first attempt at this might be to extend the Brillinger proof to imply $\mathcal{L}^2$ convergence of DFTs to the multivariate normal and use this to leverage a stronger from of convergence than convergence in distribution. Because both terms in the difference $\lim_{T\to\infty} H(\hat{X}^{(T)}) - H(\hat{X})$ likely increase to infinity with $T$ an important aspect of this puzzle will be making use of fact that we are dealing with entropy rates rather than simply entropies, as in [5].

While the Correlation Theory method for calculating the full mutual information rate is supported empirically [21], it requires not only the convergence of entropies assumption employed by our extension of the lower bound for jointly stationary processes but a second assumption also. The second assumption is that Brillinger's theorem holds for certain non-stationary stochastic processes, those that are functionals of a stationary process and a sample path from a stationary process. We spent some time using Volterra expansions [63, 2] to try to prove this second assumption but were unable to. However, we expect that such an extension of Brillinger's theorem is not outside the realm of possibility given the empirical support for it and the idea that the central limit theorem intuition behind Brillinger's work should extend to processes that are a function of a stationary sample path if that sample path is sufficiently ergodic. Extending Brillinger's theorem thus would be a worthy direction for future work.

In the original formulation of Correlation theory the expected value over sample paths that we have included (see theorem 2.5.2) is dropped without justification. It is likely that this expected value is unnecessary given the ergodicity of the sample paths employed in the theorem, but it remains to be proved. It seems likely that the same tools that one would apply to extending Brillinger's proof–cumulants, Volterra expansions, central limit theorem and ergodic properties–would prove useful for this second endeavour also.

Beyond the assumptions above, which are inherent in the use of the lower bound method to calculate the full information rate and in Correlation Theory, the information theoretic analysis employed by this study assumes that the input O-U

processes and output spike-trains produced by the $\text{SRM}_0^2$ model are jointly stationary and are all ergodic and finite memory processes. The O-U process inputs are known to satisfy the latter two properties and to be stationary themselves. This given, it seems likely that the output spike-trains of the model satisfy the necessary properties because the model itself has finite memory. Furthermore, it is shown in Brillinger's book on time series analysis [11] (using Theorems 2.9.1 and 2.3.2) that if a system can be modeled with a finite Volterra series expansion whose coefficients are absolutely summable (intuitively this should be related to the memory properties of the system) and the inputs to the system are finite memory then the outputs of the system also satisfy this property. We thus think it likely that the assumptions of joint stationarity, finite memory and ergodicity are at least approximately satisfied.

The $\text{SRM}_0^2$ model is particularly interesting as it is a self-inhibiting process and self-inhibiting processes have, to our knowledge, received little study compared with self-exciting processes (e.g. the Hawke's process [71, 52]). For the above reasons, and because the $\text{SRM}_0^2$ model provides a concise, easily parameterized computational model of a bursting neuron, a dedicated study of the properties of the $\text{SRM}_0^2$ would be an interesting direction for future study.

### 4.0.2 Implications & Future Directions in Neuroscience

While this thesis initially set out to narrow down the set of cells capable of supporting a burst multiplexing neural code, the results of our study prove perhaps more interesting from a neuroscience perspective: that burst multiplexing can perform both in unimodal ISI regimes and a range of bimodal regimes that appears to encompasses those which are physiologically likely. This finding provides strong support for the utility of a burst-multiplexing neural code in neurons across the brain, warranting future research in experimental and theoretical neuroscience.

The next step in experimental neuroscience is to check whether the brain is indeed using this multiplexing code and, if so, how. A likely candidate function for burst multiplexing is to use the burst channel to transmit an error signal, where 'error' signal is used in a supervised-learning sense, while the event channel transmits sensory information. One way of checking this hypothesis would be to design a perceptual experiment where the participant or subject must correctly identify a stimulus. In this paradigm the time series of participant errors would represent the burst input and the stimulus signal would represent the event input. If error and stimulus signals were approximately independent one could evaluate the occurrence of burst-multiplexing in the brain by applying the information rate estimators used in this study to the two input signals and the spike trains or membrane potentials recorded in the experiment.

In theoretical neuroscience, a clear future direction is to investigate whether this phenomenon is reproduced in a model where bursts are composed of more than two spikes. This is motivated by the fact that in many cells bursts contain more than

two spikes [45] and the intuition that decoding cells should be able to identify bursts more accurately when they can make use of more than one ISI to do so. A parallel line of theoretical research should address the important role of the post-synaptic synapse in decoding a burst-multiplexed signal–specifically how one might get more out of a burst multiplexing code by using more realistic weight functions than the Heaviside step function. We intend to address both of the above mentioned problems at the same time by adapting the $\mathrm{SRM}_0^2$ to allow for bursts of more than two spikes and by using an exponential function with memory of more than one ISI to represent a more flexible, biologically plausible decoding function.

### 4.0.3 Conclusion

This thesis has made contributions to theoretical neuroscience and applied math. Regarding the latter, it was shown that the Correlation Theory result of Dettner & Münzberg [21] can be applied not only to spike trains but to other types of stationary processes and a different, more efficient, proof of Correlation Theory has been given than was presented in the original work. It has also been suggested that the lower bound on the information rate, a well-used tool in the neuroscience community, may be equal to the full information rate for a larger subset of processes than simply Gaussian processes, as a consequence of Brillinger's theorem (Th. 2.3.5). Still, this result remains to be rigorously developed.

Regarding neuroscience, it has been shown that a burst multiplexing neural code is robust to Inter-Spike Interval (ISI) distribution shape, specifically that the separation of ISI mode arising from intra-burst spikes and the mode arising from all other ISIs need not be as large as one might think for effective decoding. These results will hopefully inspire new experimental research and motivate the neuroscience community to embark on a deeper study of bursting neural codes.

# Appendix A

# Computational Methods

# A.1 Computational Model

The numerical implementation of the model neural network will be discussed in this section.

## A.1.1 Input Signals

The randomness inherent in the Ornstein-Uhlenbeck process allows exact numerical generation of process sample paths, in contrast to the approximate numerical solutions that one can generate for ODEs. This method, outlined by Gillespie [32], was employed with parameters (see definition 2.1.9) given in the table below.

| Parameter | Symbol (Units) | Event Input | Burst Input |
|---|---|---|---|
| **Variance** | $\boldsymbol{\sigma^2}$ **(mV$^2$ms)** | 20 | 240 |
| **Time Constant** | $\boldsymbol{\tau}$ **(ms)** | 10 | 20 |

Table A.1: Input Signal Parameters

Larger variance was chosen for the burst channel as we found this was necessary to result in sufficient information encoding; the larger time constant for the burst stream was selected based on the rational outlined in section 3.1.

## A.1.2 Encoding Population

The following link functions were used to generate burst probability and event rate from the membrane potentials employed in the $\text{SRM}_0^2$ model (see definition 2.2.3).

$$f_{\text{link,b}}(v) = \exp\left(\frac{v-\theta_e}{\alpha_e}\right) \tag{A.1.1}$$

$$f_{\text{link,e}}(v) = \sigma\left(\frac{v-\theta_b}{\alpha_b}\right) \tag{A.1.2}$$

The exponential link function was chosen as it is commonly used in neural rate models [31]. The sigmoid link function, denoted by $\sigma$ above, was selected because it is the natural option for converting values on the positive real line to probabilities and has been used for burst modeling historically [54]. An exponential function was also used to model the relative refractory time period of the neurons, as is standard [31],

because the exponential decay reproduces the biological phenomenon appropriately. We note that a potential future direction could be to explore other relative refractory period function shapes. $t'$ is the time of the last spike.

$$\eta(t - t') = -\exp\left(-\frac{t - t'}{\tau_{rel}}\right) \tag{A.1.3}$$

The parameters for the $\text{SRM}_0^2$ model are given in the table below, followed by the rationale for parameter choices.

| Parameter Name | Symbol (Units) | Parameter Value |
|---|---|---|
| **Cell Resting Potential** | $\boldsymbol{v_0}$ **(mV)** | 0 |
| **Burst Threshold** | $\boldsymbol{\theta_b}$ **(mV)** | 4.5 |
| **Burst Scale Factor** | $\boldsymbol{\alpha_b}$ **(mV)** | $\frac{1}{3}$ |
| **Burst ISI Distribution Scale** | $\boldsymbol{\Gamma_1}$ **(ms)** | $\frac{20}{3}$ |
| **Burst ISI Distribution Shape** | $\boldsymbol{\Gamma_2}$ **(unitless)** | 1.5 |
| **Absolute Refractory Period** | $\boldsymbol{\Delta_{\text{ref}}}$ **(ms)** | 2 |
| **Relative Refractory Time Constant** | $\boldsymbol{\tau_{rel}}$ **(ms)** | Variable |
| **Event Threshold** | $\boldsymbol{\theta_e}$ **(mV)** | Variable |
| **Event Scale Factor** | $\boldsymbol{\alpha_e}$ **(mV)** | 2 |
| **Encoding Population Size** | $\boldsymbol{N}$ **(unitless)** | 200 |

Table A.2: $\text{SRM}_0^2$ Model Parameters

The cell resting potential doesn't effect information transmitted and was set to zero for simplicity. The burst threshold and scale factor were chosen so that the fraction of total events that are bursts was approximately 0.2, as the burst fraction measured in layer 2/3 and layer 5 cortical cells is measured to be in the range 0.1 to 0.2 [19]. The burst ISI scale parameter was chosen so that rate of spikes generated by a sequence of burst ISIs would be in the 100-200 Hz range observed in layer 5 cells [44] and the shape parameter was selected to qualitatively produce the sharp, super-exponential peak observed in ISI distributions [6]. The 2 ms absolute refractory period is in line with the literature on cortical cells [37] as well as cells in sub-cortical regions [7].

The primary variable in the model was the relative refractory period of the event spike ISIs, $\tau_{rel}$, which was varied in 4 ms increments between 6 and 50 ms to parameterize a range from unimodal to well separated bimodal ISI distributions. Firing rate in cortical cells covers roughly two orders of magnitude, from around

1Hz to tens of Hz [61]. Accordingly, the event threshold and scale factor were set to produce an event rate of approximately 10 Hz when $\tau_{rel}$ was set to 6 ms, which led to a value of 3.29 for the event threshold. For the uncorrected rate results (figure 3.3) these initial event threshold and scale factor values remained fixed as $\tau_{rel}$ increased while for the rate corrected results (figures 3.1 and 3.2) the event threshold was decreased with increasing $\tau_{rel}$ to keep event rate fixed. Lastly, 200 $\mathrm{SRM}_0^2$ cells were used for the encoding population as the goal was to explore a regime where spike generation (finite size effect) noise would be appreciable.

### A.1.3 Decoding Cells

The decoding cells were composed of synaptic weight functions (see section 2.2.2), a synaptic filter $\kappa_{syn}(t)$ and a resting synaptic potential. The weight functions were described by a single threshold parameter for burst, $\tau_{\mathrm{w},b}$, and event $\tau_{\mathrm{w},e}$ decoding cells respectively. In all simulations these were set to the same value: $\tau_{\mathrm{w}}$. The synaptic filter was modeled as an exponential rise and exponential decay as this double exponential response to incoming spikes is typically observed at neural synapses [46]. Notably, the information rate itself would not be effected by the shape of the kernel used because this linear filter will factor out of the expressions for information rate.

$$\kappa_{syn}(t) = \bar{g}\left(1 - \exp\left(\frac{-t}{\tau_{\mathrm{rise}}}\right)\right)\exp\left(\frac{-t}{\tau_{\mathrm{decay}}}\right) \tag{A.1.4}$$

The parameters for decoding cells are listed below, followed by a motivation of their choice.

| Parameter Name | Symbol (Units) | Parameter Value |
|---|---|---|
| **Resting Synaptic Potential** | $\boldsymbol{v_{syn}}$ **(mV)** | 1 |
| **Synaptic Filter Scale** | $\bar{\boldsymbol{g}}$ **(mV)** | 1 |
| **Synaptic Rise Constant** | $\boldsymbol{\tau_{\mathrm{rise}}}$ **(ms)** | 3 |
| **Synaptic Decay Constant** | $\boldsymbol{\tau_{\mathrm{decay}}}$ **(ms)** | 5 |
| **Weight Function Threshold** | $\boldsymbol{\tau_{\mathrm{w}}}$ **(ms)** | Variable |

Table A.3: Parameters for Decoding Cells

The important aspect of the decoding cell parameterization was to arrive at quantities which were equivalent, in an information theoretic sense, to the synapses of decoding cells. For this reason their values were somewhat arbitrary, baring the

considerations mentioned below. The resting synaptic potential does not effect information transmission but, if set to zero, it would result in an undefined estimate of burst fraction in a period of prolonged encoding population inactivity. The synaptic filter scale doesn't effect information and was set for simplicity. Because the synaptic filter itself is linear it will not effect information and was included simply for completeness. The weight function threshold was chosen using a simple grid search to maximize information (see figure 3.1), as discussed in section 3.1.

## A.2 Simulation and Data Analysis

The model network described in the above section was simulated using python 3.6 with a discretization of 1 ms because it allowed for a measurement of frequencies in the range $[0, 500]$ Hz, and we noted that information dropped to negligible amounts well before this threshold (i.e. even before 200 Hz).

### A.2.1 Linear Lower Bound

To calculate information rate using the linear lower bound method (Th. 2.5.1) the network was simulated for 819.200 seconds and Welch's method [78] with a Hanning window [34], window length of 8.192 seconds and an overlap of 4.096 seconds was used to calculate the coherence. The integral over frequency was solved using Simpson's rule. For the burst channel this numerical integration was over the range $[0, 75]$ Hz as negligible amounts of information were present above this bandwidth. Equivalently, for the event channel, integration was over $[0, 150]$ Hz.

### A.2.2 Correlation Theory

For figure 3.2 the variances appearing in equation 2.5.3 were calculated by simulating the model for 600 trials of 49.152 seconds, dividing the trials into periods of 16.384 second sub-trials, using the Fast Fourier Transform (FFT) (see Appendix C.2) to calculate the DFT at each frequency for each of the 1800 sub-trials and then calculating the variance of the DFT at each frequency from the set of sub-trials. This resulted in one estimate of the expression inside the expected value in equation 2.5.3. The expected value was implemented by calculating the sample mean over 7 estimates. Only 7 were used because there was not much variance in the estimates. The integral over frequency was implemented in exactly the same fashion as for the linear lower bound.

For figure 3.3 the same process was used but with 500 trials and averaging over 12 stimuli for the expected value.

# Appendix B

# Numerical Validation

## B.1 Rate Correction

This section validates that rate was properly corrected for in figures 3.2 and C.1. Expected value of $\text{SRM}_0^2$ rate could not be calculated analytically so a numerical approach was taken. Spike counts (numerator of rate) were adjusted so that linear regression with spike counts as $y$ variable and $\tau_{rel}$ as $x$ variable produced a non-significant slope. To verify visually, the normalized, summed spike counts generated from the ISIs associated with figure 3.2 are plotted below. Note the small range of the $y$-axis.

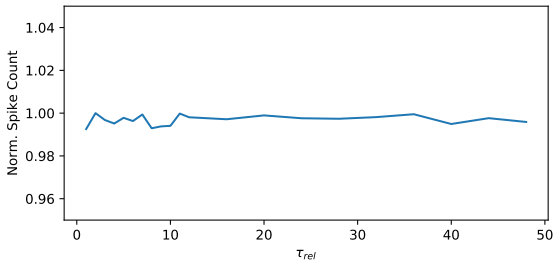


Figure B.1: **Validation of Rate Correction:** $Y$-axis is the total spike counts per ISI distribution, normalized by the maximum of these over parameter values. $X$-axis is parameter value

## B.2 Validation of Correlation Theory Assumptions

This section demonstrates that the properties of Brillinger's proof hold for the non-stationary process that is the output of the event (burst) channel given the input to the event (burst) channel. Figures B.2 and B.3 give asymptotic normality and independence, respectively, for the outputs of the event (burst) channel given a fixed sample path of the input signal to that channel. Figures B.4 and B.5 are the same as B.2 and B.3 but without fixing the sample path, numerically confirming Brillinger's theorem (assuming the outputs of the $\text{SRM}_0^2$ are indeed stationary), for comparison with the non-stationary case.

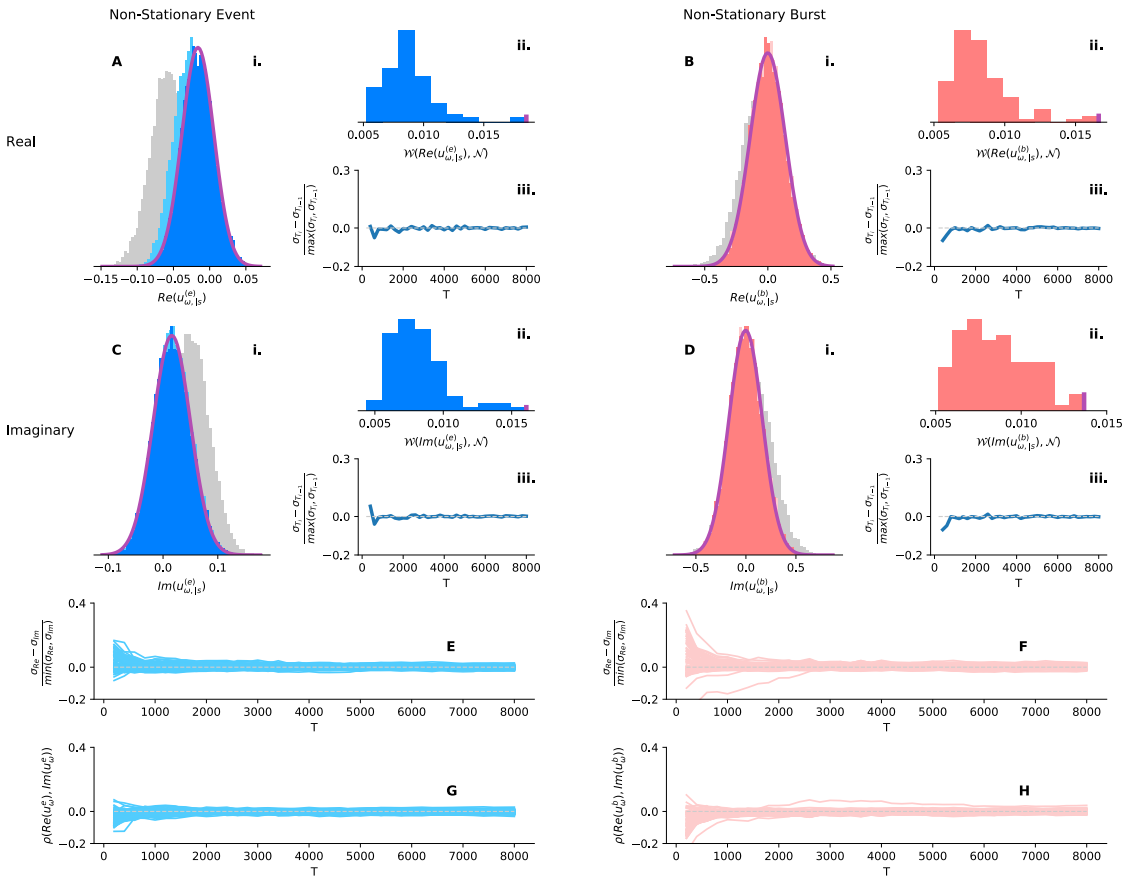


Figure B.2: **Asymptotic Normality, Non-Stationary:** First row is for real part, second is imaginary. Blue denotes event channel and orange burst. **A-D: i.** Distribution of DFT real or imaginary part at one frequency. Purple line is fitted normal. Shadow distributions are DFTs generated from shorter time series; note convergence to darkest colour distribution (which is generated by a 8192 ms length time series). **ii.** Wasserstein (see Appendix C.3) distance between distribution of DFT at one frequency and normal distribution. The histogram is over a sample of DFT frequencies. Vertical purple line denotes point in histogram corresponding to the DFT from i. (meaning all other samples fit the normal distribution even better). **iii.** Convergence of standard deviation of DFT part from i., as length of time series used to calculate DFT increases . Plotted value is difference between standard deviations of DFTs calculated from $T$ and $T-1$ length time series, normalized by the maximum value of the two standard deviations. **E, F:** Convergence of variance of real and imaginary parts of DFT at one frequency to the same value. Y-axis is the difference in variance of the real and imaginary parts normalized by the minimum of the two variances, x-axis is length of time series used to calculate DFT. **G, H:** Asymptotic independence of real and imaginary components at one frequency. Y-axis is the correlation of real and imaginary parts; x-axis is length of time series used in DFT calculation. Grey dashed line denotes 0.

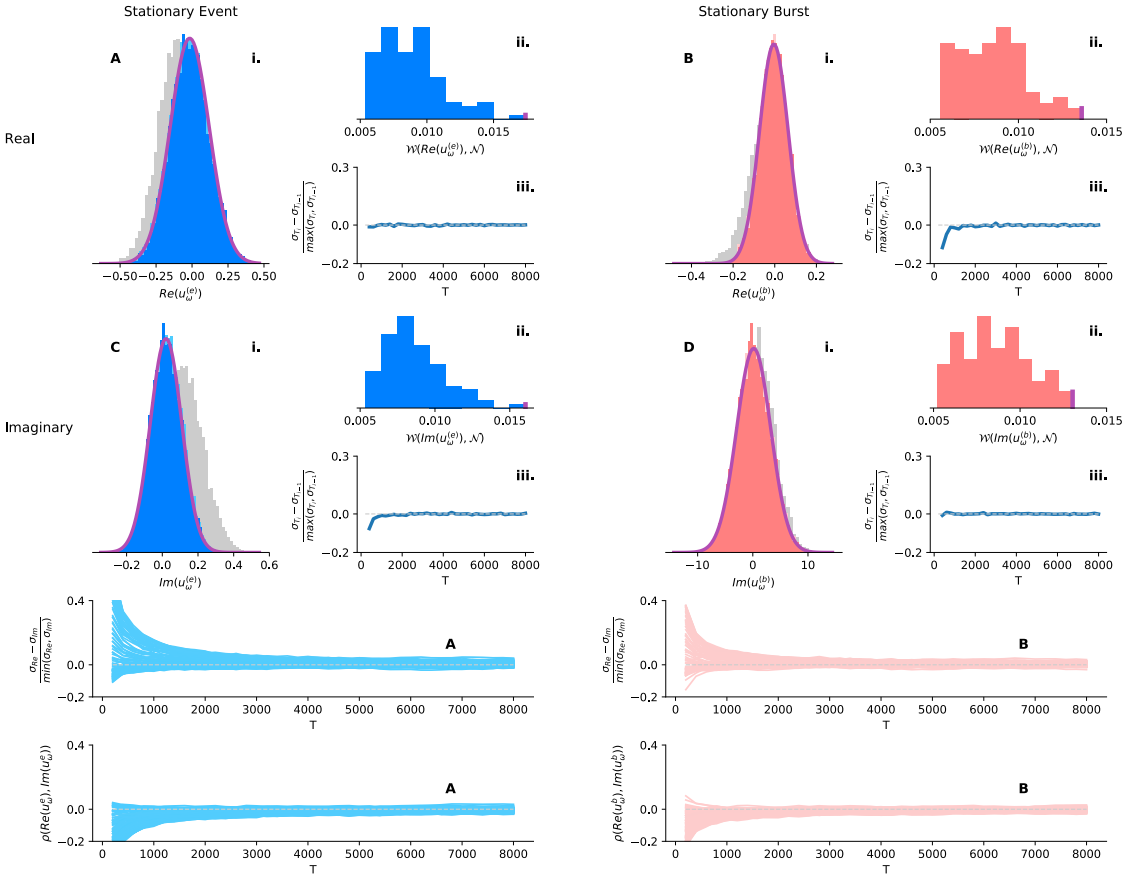


Figure B.3: **Asymptotic Normality, Stationary:** See figure B.2 caption for description.

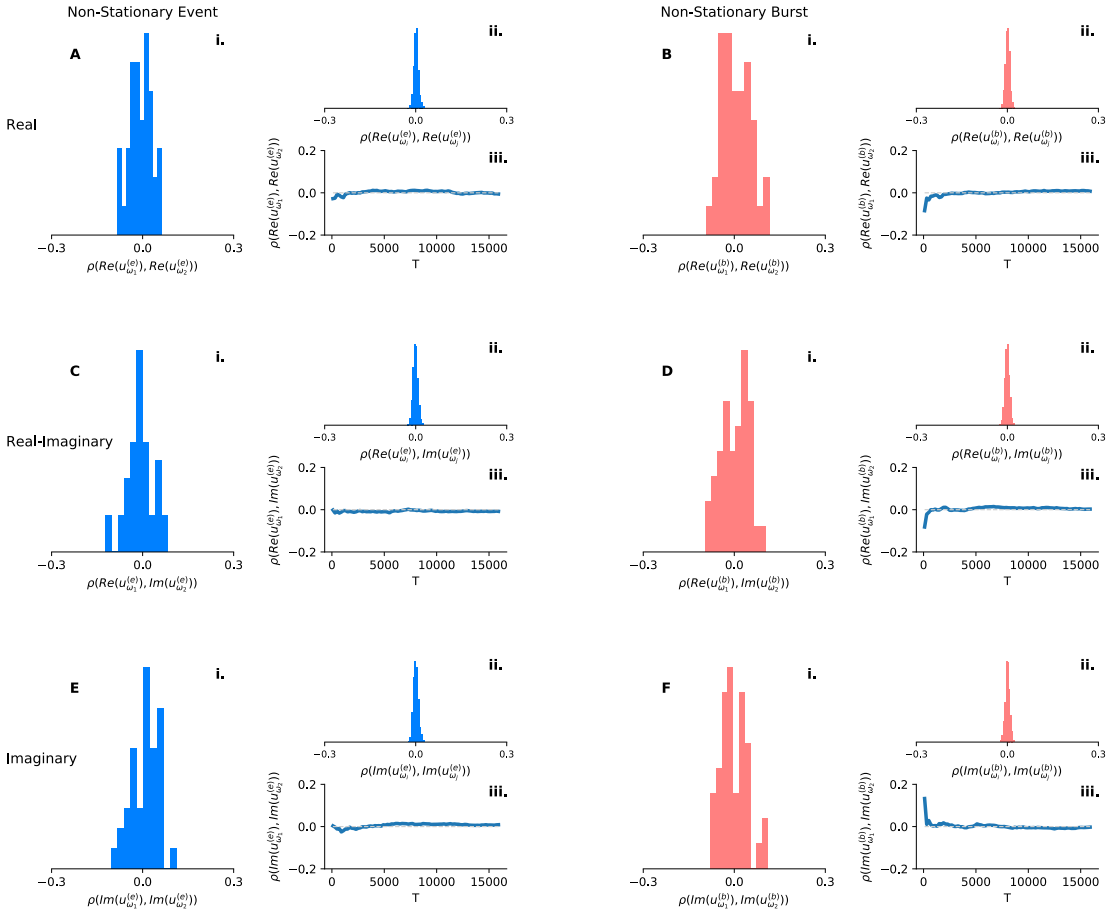


Figure B.4: **Asymptotic Independence, Non-Stationary:** First row is for correlation between pairs of real parts, second is correlation between pairs of real and imaginary parts, third is correlation between pairs of imaginary parts; all pairs are of DFTs at different frequencies. Blue denotes event channel and orange burst. **A-F: i.** Histogram of correlation coefficients of one pair of DFT frequencies. This histogram has a wider spread than the one in ii. because each correlation coefficient is calculated from a smaller sample size than those in ii. Note that distribution is centered at zero. DFT calculated from 8192 ms length time series. **ii.** Frequency components are uncorrelated: plotted is a histogram of correlation coefficients of a sample of pairs of DFTs at different frequencies. DFT calculated from 8192 ms length time series. **iii.** Asymptotic independence of pair of frequencies: Y-axis is correlation coefficient of the pair of DFTs pictured in i. and x-axis is length of time series used to calculate DFTs. Note rapid convergence to zero.

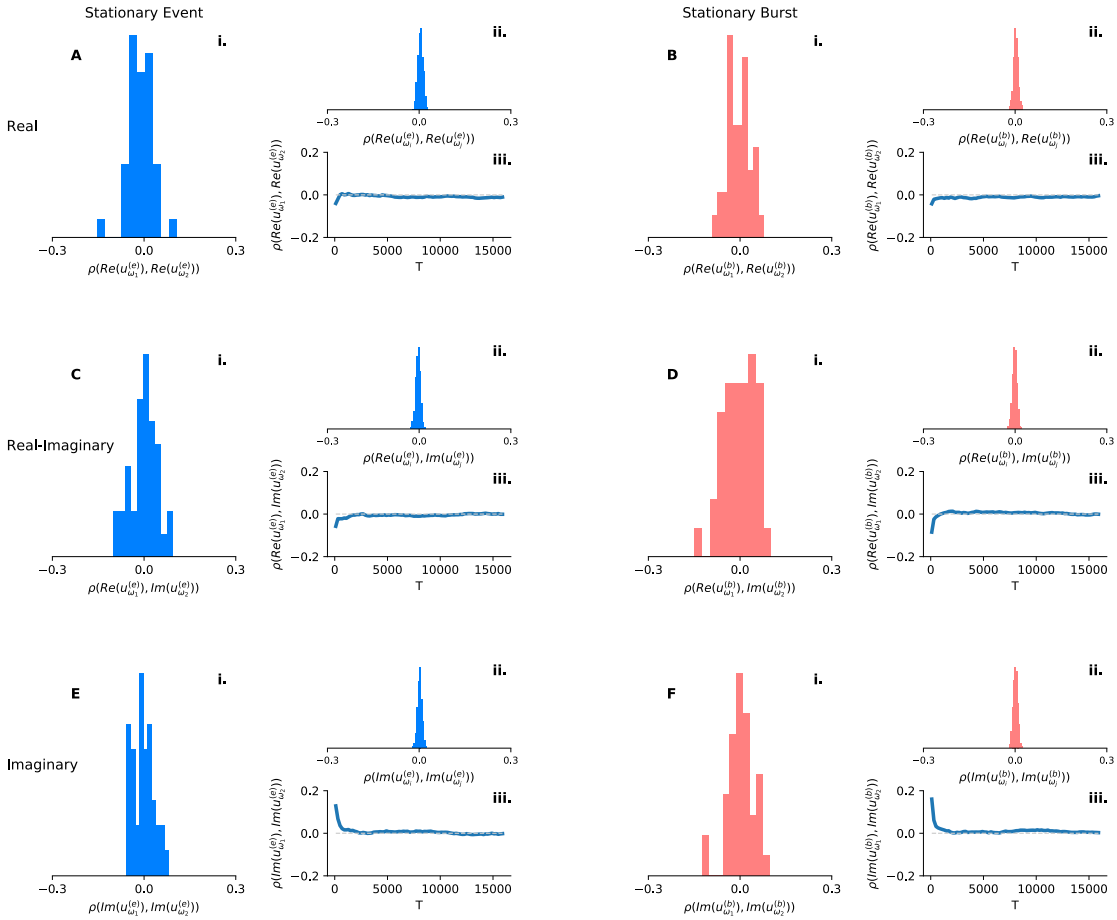


Figure B.5: **Asymptotic Independence, Stationary:** See figure B.4 caption for description.

# Appendix C

# Miscellaneous

# C.1 Complex Exponentials

This section demonstrates two properties of complex exponentials: (1) the shift property, that leads to one being able to shift back and forth between an index centered at zero or beginning at zero for the DFT without changing the magnitude of the DFT; (2) the orthogonality of vectors of complex exponentials.

## C.1.1 Shift

Let $n \in \{-N+\alpha, -N+1+\alpha, ..., N-1, N\}$ be a $T = 2N$ length index sequence to which the DFT is being applied and let $\alpha = 0$ or 1 according to whether $N$ is odd or even respectively. Let $x$ be arbitrary. Then

$$\sum_{n=-N+\alpha}^{N} \exp(inx) = \sum_{n=-N+\alpha}^{N} \exp\big(i(n+N-\alpha-N+\alpha)x\big) \tag{C.1.1}$$

Define $k = n + N - \alpha$. Then

$$= \sum_{k=0}^{2N-\alpha} \exp(ikx) \exp\big(-i(N+\alpha)x\big) = c \sum_{n=-N+\alpha}^{N} \exp(ikx) \tag{C.1.2}$$

where $|c| = 1$. It follows that if one is dealing with a shift invariant sequence (e.g. a stationary stochastic process), changing the index of a DFT only multiplies the DFT by a factor of magnitude one, thus leaving the power spectrum unchanged.

## C.1.2 Orthogonality

The concept at the heart of the asymptotic independence of Fourier components is the asymptotic orthogonality of certain vectors of complex exponentials. Let $X = [e^{\frac{2\pi 0x}{N}}\ e^{\frac{2\pi 1x}{N}}\ ...\ e^{\frac{2\pi(N-1)x}{N}}\ e^{\frac{2\pi Nx}{N}}]^T$ and $Y = [e^{\frac{2\pi 0y}{N}}\ e^{\frac{2\pi 1y}{N}}\ ...\ e^{\frac{2\pi(N-1)y}{N}}\ e^{\frac{2\pi Ny}{N}}]^T$. Let $x, y \in \mathbb{Z}$ such that $x, y \leq N$ and $x \neq y$. Now consider the inner product of $X$ and $Y$

$$\sum_{n=0}^{N-1} X_n Y_n = \sum_{n=0}^{N-1} e^{\frac{2\pi n(x-y)}{N}} = \frac{1-e^{2\pi(x-y)}}{1-e^{\frac{2\pi(x-y)}{N}}} = \frac{1-1}{1-e^{\frac{2\pi(x-y)}{N}}} = 0 \tag{C.1.3}$$

where for the second equality a geometric progression is used. Thus if $x \neq y$ then $X$ and $Y$ are orthogonal. Conversely, if $x = y$ then the geometric progression is undefined but each term of the sum becomes equal to 1 so the entire sum equals $N$. Equation 2.3.14 follows from these observations.

## C.2 DFT Frequency Order

From a finite $2N + \alpha = T \in \mathbb{N}$, where $\alpha$ is 0 or 1 according to whether $T$ is even or odd, length sequence of time samples one can calculate the DFT at as many different frequencies as one likes. However, the DFT is a periodic function of frequency, with a period of $2\pi$, 1 or $T$ depending on the frequency units one adopts. For this reason, to identity the frequency domain function represented by the DFT one need only calculate frequencies within a single period. The orthogonal DFT, which is a basis transformation and is used throughout the thesis, obviously requires the orthogonality property outlined in the previous section and thus calculates DFT values at frequencies that are rational numbers $\frac{k}{T}$, where $|k| \leq T$ (or $k \in \mathbb{Z}$ such that $0 \leq k < T$, if one shifts the index to begin at zero).

Define frequencies to be the values $f$ such that $f = \frac{r}{T}$ s.t. $r \in \{-N + \alpha, -N + 1 + \alpha, ..., 0, 1, 2, ..., N - 1, N\}$ (i.e. $f$ is defined as in the proof of Brillinger's theorem, Th. 2.3.5, but labeling the index variable as $k$ instead of $r$ and allowing even length sequences via $\alpha$). Now define $n = r + N - \alpha$ and use this variable change to shift the DFT so that the index begins at zero. **n defines the index used for the frequencies in lemma 2.4.23**. By the shift property mentioned in the previous section this will only multiply the DFT by a constant of magnitude one and should thus effect neither mutual information nor entropy.

## C.3 Wasserstein Distance

The Wasserstein distance gives a measure of divergence between probability distributions analogously to the KL divergence discussed in section 2.4. Unlike the KL divergence however, it is a metric (because, in contrast to KL divergence, it satisfies the symmetry property). The $k^{th}$ Wasserstein distance is defined below

$$W_k(p, q) = \left( \inf_{\nu \in \mu} \mathrm{E}_\nu \left[ d(X, Y)^k \right] \right)^{1/k} \tag{C.3.1}$$

where $(\Omega, d)$ is the metric space, $k \in \mathbb{N}$, $p$ and $q$ are probability measures on $\Omega$, and $\mu$ is the set of all probability measures on $\Omega \times \Omega$ with marginals $p$ and $q$. In this thesis the first Wasserstein distance was used with $d(x, y) = |x - y|$. It is used in Appendix B.2 to show similarity between the distribution of DFT real or imaginary parts at a given frequency and the normal distribution. Specifically, the standard normal and the centered DFT distribution with variance normalized to one were compared.

## C.4 Primary Results via Lower Bound

Below is figure 3.2 reproduced using the lower bound method (Th. 2.5.1) rather than Correlation theory (Th. 2.5.2).

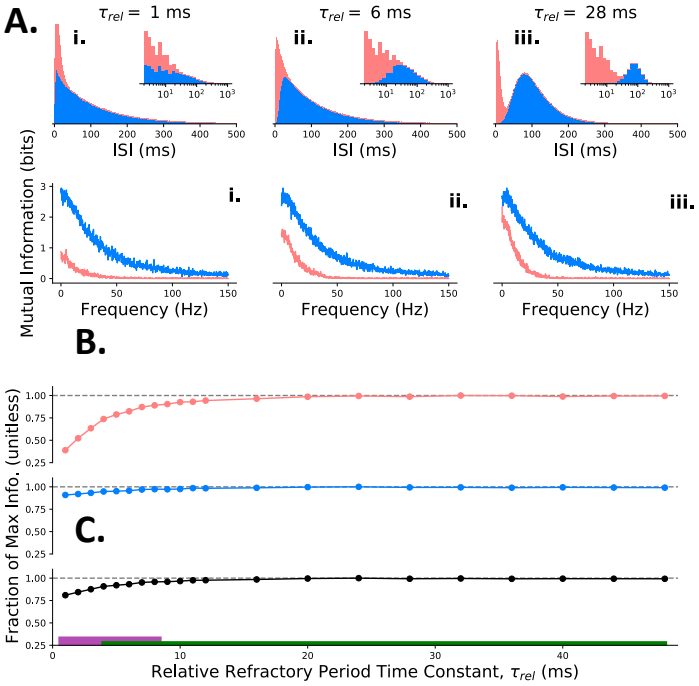


Figure C.1: **Burst Multiplexing is Robust to ISI Shape, Lower Bound Method:** Mutual information rate is minimally effected by ISI distribution shape. **A.** ISI distributions for three values of relative refractory period time constant: **i.** 1 ms, **ii.** 6 ms, **iii.** 28 ms **B.** Mutual information as a function of frequency for encoding populations with the ISI distributions in (A). **C.** Fraction of maximum information rate (max taken over all tested ISI distribution shapes) as a function of relative refractory period time constant (ISI distribution shape) in **i.** the burst channel (orange) **ii.** the event channel (blue) **iii.** the sum of channels. Dashed line indicates ratio of 1, green region (at bottom) denotes parameter range over which more than 70% of the max information is transmitted and purple region denotes parameter range for unimodal ISI distributions; note the overlap.

# Bibliography


[1] Edgar D Adrian and Yngve Zotterman. The impulses produced by sensory nerve-endings: Part ii. the response of a single end-organ. *The Journal of physiology*, 61(2):151–171, 1926.

[2] Nasir Uddin Ahmed. *Generalized Functionals of Brownian Motion and Their Applications: Nonlinear Functionals of Fundamental Stochastic Processes*. World Scientific, 2012.

[3] Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha. Efficient machine learning for big data: A review. *Big Data Research*, 2(3):87–93, 2015.

[4] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358, 2006.

[5] Andrew R Barron et al. The strong ergodic theorem for densities: generalized shannon-mcmillan-breiman theorem. *The annals of Probability*, 13(4):1292–1303, 1985.

[6] Joseph Bastian and Jerry Nguyenkim. Dendritic modulation of burst-like firing in sensory neurons. *Journal of Neurophysiology*, 85(1):10–22, 2001.

[7] Michael J Berry II and Markus Meister. Refractoriness and neural precision. In *Advances in Neural Information Processing Systems*, pages 110–116, 1998.

[8] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[9] Romain Brette. Philosophy of the spike: rate-based vs. spike-based theories of the brain. *Frontiers in systems neuroscience*, 9:151, 2015.

[10] William L Briggs et al. *The DFT: an owners' manual for the discrete Fourier transform*, volume 45. Siam, 1995.

[11] David R Brillinger. *Time series: data analysis and theory*, volume 36. Siam, 1981.

[12] Nicolas Brunel and Mark CW Van Rossum. Lapicque's 1907 paper: from frogs to integrate-and-fire. *Biological cybernetics*, 97(5-6):337–339, 2007.

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

[14] A Cattaneo, L Maffei, and Concetta Morrone. Patterns in the discharge of simple and complex visual cortical cells. *Proceedings of the Royal Society of London. Series B. Biological sciences*, 212(1188):279–297, 1981.

[15] Albert Compte, Christos Constantinidis, Jesper Tegnér, Sridhar Raghavachari, Matthew V Chafee, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *Journal of neurophysiology*, 90(5):3441–3454, 2003.

[16] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[17] David R Cox. The statistical analysis of dependencies in point processes. *Stochastic point processes*, 55:66, 1972.

[18] Dorota M Dabrowska and Wai Tung Ho. Estimation in a semiparametric modulated renewal process. *Statistica Sinica*, pages 93–119, 2006.

[19] CPJ De Kock and Bert Sakmann. High frequency action potential bursts ($\geq$ 100 hz) in l2/3 and l5b thick tufted neurons in anaesthetized and awake rat primary somatosensory cortex. *The Journal of physiology*, 586(14):3353–3364, 2008.

[20] Alain Destexhe, Michael Rudolph, J-M Fellous, and Terrence J Sejnowski. Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons. *Neuroscience*, 107(1):13–24, 2001.

[21] Amadeus Dettner, Sabrina Münzberg, and Tatjana Tchumatchenko. Temporal pairwise spike correlations fully capture single-neuron information. *Nature communications*, 7:13805, 2016.

[22] Peter U Diehl, Guido Zarrella, Andrew Cassidy, Bruno U Pedroni, and Emre Neftci. Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8. IEEE, 2016.

[23] Brent Doiron, André Longtin, Neil Berman, and Leonard Maler. Subtractive and divisive inhibition: effect of voltage-dependent inhibitory conductances and noise. *Neural computation*, 13(1):227–248, 2001.

[24] Joseph L Doob. The brownian movement and stochastic equations. *Annals of Mathematics*, pages 351–369, 1942.

[25] Iddo Eliazar and Joseph Klafter. On the nonlinear modeling of shot noise. *Proceedings of the National Academy of Sciences*, 102(39):13779–13782, 2005.

[26] Jan Eriksson, Esa Ollila, and Visa Koivunen. Statistics for complex random variables revisited. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3565–3568. IEEE, 2009.

[27] Stanley Finger. *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, USA, 2001.

[28] Fabrizio Gabbiani. Coding of time-varying signals in spike trains of linear and half-wave rectifying neurons. *Network: Computation in Neural Systems*, 7(1):61–85, 1996.

[29] Michael A Gaffield, Audrey Bonnan, and Jason M Christie. Conversion of graded presynaptic climbing fiber activity into graded postsynaptic ca2+ signals by purkinje cell dendrites. *Neuron*, 102(4):762–769, 2019.

[30] Wulfram Gerstner. Time structure of the activity in neural network models. *Physical review E*, 51(1):738, 1995.

[31] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[32] Daniel T Gillespie. Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Physical review E*, 54(2):2084, 1996.

[33] Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[34] Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

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

[36] Martin Jacobsen. *Point process theory and applications: marked point and piecewise deterministic processes*. Springer Science & Business Media, 2006.

[37] Prakash Kara, Pamela Reinagel, and R Clay Reid. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron*, 27(3):635–646, 2000.

[38] Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*, 13(8):1713–1720, 2001.

[39] JFC Kingman. Poisson processes.— clarendon press, 1993.

[40] Robert J Kosinski. A literature review on reaction time. *Clemson University*, 10, 2008.

[41] Rüdiger Krahe and Fabrizio Gabbiani. Burst firing in sensory systems. *Nature Reviews Neuroscience*, 5(1):13, 2004.

[42] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[43] Y Lamarre, M Filion, and JP Cordeau. Neuronal discharges of the ventrolateral nucleus of the thalamus during sleep and wakefulness in the cat i. spontaneous activity. *Experimental brain research*, 12(5):480–498, 1971.

[44] Matthew E Larkum, KMM Kaiser, and Bert Sakmann. Calcium electrogenesis in distal apical dendrites of layer 5 pyramidal cells at a critical frequency of back-propagating action potentials. *Proceedings of the National Academy of Sciences*, 96(25):14600–14604, 1999.

[45] Matthew E Larkum, Thomas Nevian, Maya Sandler, Alon Polsky, and Jackie Schiller. Synaptic integration in tuft dendrites of layer 5 pyramidal neurons: a new unifying principle. *Science*, 325(5941):756–760, 2009.

[46] Matthew Evan Larkum, Jack Waters, Bert Sakmann, and Fritjof Helmchen. Dendritic spikes in apical dendrites of neocortical layer 2/3 pyramidal neurons. *Journal of Neuroscience*, 27(34):8999–9008, 2007.

[47] Nicholas A Lesica and Garrett B Stanley. Encoding of natural scene movies by tonic and burst spikes in the lateral geniculate nucleus. *Journal of Neuroscience*, 24(47):10731–10740, 2004.

[48] Stefan Leutgeb, Jill K Leutgeb, Alessandro Treves, May-Britt Moser, and Edvard I Moser. Distinct ensemble codes in hippocampal areas ca3 and ca1. *Science*, 305(5688):1295–1298, 2004.

[49] Michael J Lighthill, Michael James Lighthill, et al. *An introduction to Fourier analysis and generalised functions*. Cambridge University Press, 1958.

[50] John E Lisman. Bursts as a unit of neural information: making unreliable synapses reliable. *Trends in neurosciences*, 20(1):38–43, 1997.

[51] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[52] Maxime Morariu-Patrichi and Mikko S Pakkanen. Hybrid marked point processes: characterisation, existence and uniqueness. *arXiv preprint arXiv:1707.06970*, 2017.

[53] Hiroshi Nakahama, Hisao Suzuki, Mitsuaki Yamamoto, Sadao Aikawa, and Shinko Nishioka. A statistical analysis of spontaneous activity of central single neurons. *Physiology & Behavior*, 3(5):745–752, 1968.

[54] Richard Naud and Henning Sprekeler. Sparse bursts optimize information transmission in a multiplexed neural code. *Proceedings of the National Academy of Sciences*, 115(27):E6329–E6338, 2018.

[55] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, page 1, 2018.

[56] Athanasios Papoulis and S Unnikrishna Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[57] Francisco J Piera and Patricio Parada. On convergence properties of shannon entropy. *Problems of Information Transmission*, 45(2):75–94, 2009.

[58] H Vincent Poor. *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

[59] Joy Putney, Rachel Conn, and Simon Sponberg. Timing is (almost) everything in a comprehensive, spike-resolved flight motor program. *bioRxiv*, page 602961, 2019.

[60] Fred Rieke, David Warland, Rob de Ruyter Van Steveninck, William S Bialek, et al. *Spikes: exploring the neural code*, volume 7. MIT press Cambridge, 1999.

[61] Alex Roxin, Nicolas Brunel, David Hansel, Gianluigi Mongillo, and Carl van Vreeswijk. On the distribution of firing rates in networks of cortical neurons. *Journal of Neuroscience*, 31(45):16217–16226, 2011.

[62] Terence D Sanger. Neural population codes. *Current opinion in neurobiology*, 13(2):238–249, 2003.

[63] Martin Schetzen. *The Volterra and Wiener theories of nonlinear systems*, volume 1. Wiley New York, 1980.

[64] Volker Schmidt. *Stochastic geometry, spatial statistics and random fields*. Springer, 2014.

[65] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[66] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[67] Shigeru Shinomoto, Keiji Miura, and Shinsuke Koyama. A measure of local variation of inter-spike intervals. *Biosystems*, 79(1-3):67–72, 2005.

[68] Philip L Smith. From poisson shot noise to the integrated ornstein–uhlenbeck process: Neurally principled models of information accumulation in decision-making and response time. *Journal of Mathematical Psychology*, 54(2):266–283, 2010.

[69] William R Softky and Christof Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *Journal of Neuroscience*, 13(1):334–350, 1993.

[70] Richard B Stein, Andrew S French, and Andrew V Holden. The frequency response, coherence, and information capacity of two neuronal models. *Biophysical journal*, 12(3):295–322, 1972.

[71] Neta Ravid Tannenbaum and Yoram Burak. Theory of nonstationary hawkes processes. *Physical Review E*, 96(6):062314, 2017.

[72] Tom Tetzlaff, Moritz Helias, Gaute T Einevoll, and Markus Diesmann. Decorrelation of neural-network activity by inhibitory feedback. *PLoS computational biology*, 8(8):e1002596, 2012.

[73] Yoshiaki Tsunoda and Shinji Kakei. Reaction time changes with the hazard rate for a behaviorally relevant event when monkeys perform a delayed wrist movement task. *Neuroscience letters*, 433(2):152–157, 2008.

[74] Charles Van Loan. *Computational frameworks for the fast Fourier transform*, volume 10. Siam, 1992.

[75] Jakob Voigts and Mark T Harnett. Somatic and dendritic encoding of spatial variables in retrosplenial cortex differs during 2d navigation. *Neuron*, 2019.

[76] Benjamin Voloh, Taufik A Valiante, Stefan Everling, and Thilo Womelsdorf. Theta–gamma coordination between anterior cingulate and prefrontal cortex indexes correct attention shifts. *Proceedings of the National Academy of Sciences*, 112(27):8457–8462, 2015.

[77] Jack Waters, Matthew Larkum, Bert Sakmann, and Fritjof Helmchen. Supralinear ca2+ influx into dendritic tufts of layer 2/3 neocortical pyramidal neurons in vitro and in vivo. *Journal of Neuroscience*, 23(24):8558–8567, 2003.

[78] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[79] Jason Wolfe, Arthur R Houweling, and Michael Brecht. Sparse and powerful cortical spikes. *Current opinion in neurobiology*, 20(3):306–312, 2010.