2018

Effect of Neuromodulation of Short-Term Plasticity on Information Processing in Hippocampal Interneuron Synapses

Elham Bayat Mokhtari
University of Montana

Effect of Neuromodulation of Short-Term Plasticity on Information Processing in Hippocampal Interneuron Synapses

By

Elham Bayat Mokhtari

B.S., Applied Statistics, Ferdowsi University of Mashhad (FUM), Iran
M.S., Mathematical Statistics, Ferdowsi University of Mashhad (FUM), Iran

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctorate of Philosophy
in Mathematics

The University of Montana
Missoula, MT

December 2018

Approved by:

Scott Whittenburg, Associate Dean of the Graduate School
Graduate School

Dr. Emily F. Stone, Chair
Mathematical Sciences

Dr. Dave Patterson
Mathematical Sciences

Dr. Brian Steele
Mathematical Sciences

Dr. Johnathan Bardsley
Mathematical Sciences

Dr. Nathan Insel
Department of Psychology

"The brain is wider than the sky,
For, put them side by side,
The one the other will include
With ease, and you beside.

The brain is deeper than the sea,
For, hold them, blue to blue,
The one the other will absorb,
As sponges, buckets do.

The brain is just the weight of God,
For, lift them, pound for pound,
And they will differ, if they do,
As syllable from sound".

Emily Dickinson, 1862

Abstract

Bayat Mokhtari, Elham, Doctorate of Philosophy, December 2018 Mathematics

Effect of Neuromodulation of Short-Term Plasticity on Information Processing in Hippocampal Interneuron Synapses
Committee Chair: Emily F. Stone, Ph.D.

Neurons convey information about the complex dynamic environment in the form of signals. Computational neuroscience provides a theoretical foundation toward enhancing our understanding of nervous system. The aim of this dissertation is to present techniques to study the brain and how it processes information in particular neurons in hippocampus.

We begin with a brief review of the history of neuroscience and biological background of basic neurons. To appreciate the importance of information theory, familiarity with the information theoretic basics is required, these basics are presented in Chapter 2. In Chapter 3, we use information theory to estimate the amount of information postsynaptic responses carry about the preceding temporal activity of hippocampal interneuron synapses and estimate the amount of synaptic memory. In Chapter 4, we infer parsimonious approximation of the data through analytical expression for calcium concentration and postsynaptic response distribution when calcium decay time is significantly smaller that the interspike intervals.

In Chapter 5, we focus on the study and use of Causal State Splitting Reconstruction (CSSR) algorithm to capture the structure of the postsynaptic responses. The CSSR algorithm captures patterns in the data by building a machine in the form of visible Markov Models. One of the main advantages of CSSR with respect to Markov Models is that it builds states containing more than one histories, so the obtained machines are smaller than the equivalent Markov Model.

Acknowledgments

Foremost, I would like to express my appreciation to my committee chair Dr. Emily Stone for the continuous support of my Ph.D study and research, for her patience, motivation, and knowledge. I have been extremely lucky to have an advisor who cared about my research, and who responded to my questions and queries so promptly.

I would like to extend my gratitude to my committee members: Dr. Steele, Dr. Patterson, Dr. Bardsley, and Dr. Insel for their encouragement and invaluable support. Their insightful comments and suggestions allowed me to be more precise in theoretical and conceptual aspects of this work.

My sincere thanks also goes to Dr. Smirnova for helping me to receive the summer internship opportunity and leading me working on diverse exciting projects.

Completing this work would have been more difficult without the support provided by the members of the math department. I am indebted to them for their help.

Contents

Poem ii

Abstract iii

Acknowledgments iv

List of Figures ix

1 Introduction 1

1.1 A Brief History of Neuroscience 1

1.2 Biological Background of Basic Neurons 2

1.2.1 Neurons 2

1.2.2 Axons 3

1.2.3 Biology of Synapses 3

1.2.4 Neural Coding Theory 7

1.3 Tools for Neuroscience 8

1.3.1 Information Theory 8

1.3.2 Computational Mechanics 9

2 Information Theory **10**

2.1 Information theory in neuroscience 10

2.2 Information-Theoretic Functionals 11

2.2.1 Probability Distributions and Densities 11

2.2.2 Entropy 14

2.2.3 Joint Entropy 14

2.2.4 Conditional Entropy 15

2.2.5 Mutual Information 16

2.2.6 Differential Entropy 17

2.3 Estimation of Mutual Information 19

2.3.1 Partition Based Estimators 20

2.3.2 Metric Based Estimators 24

3 Information Processing in Hippocampal Interneuron Synapses **31**

3.1 Motivation 32

3.2 FD model 33

3.2.1 Parameter estimation 36

3.2.2 Discussion of the model 37

3.3 Numerical Study of the Response (P_r) Distributions 37

3.4 Entropy of P_r vs. Mean Firing Rate . . . 43

3.5 The Stochastic Model of the Postsynaptic Response . . . 46

3.5.1 Model description . . . 46

3.5.2 Mutual information calculations . . . 48

3.6 Information "Stored" in the Postsynaptic Response . . . 49

3.6.1 Mutual information between normalized postsynaptic response PSR and m-tuple inter-spike intervals $ISI_1, ISI_2, \cdots, ISI_m$ . . . 52

3.7 Discussion . . . 57

4 Parsimonious Approximate Descriptions of the Data 60

4.1 Motivation . . . 60

4.2 Calcium Concentration Model . . . 61

4.2.1 Characterization of the Calcium Distribution From a Random Difference Equation . . . 61

4.2.2 Random Difference Equation for Calcium Concentration . . . 63

4.2.3 Assessing the fit of the calcium concentration distribution with the Kolmogorov-Smirnov (k-s) test . . . 69

4.2.4 Quantile Gamma Graph Plot . . . 72

4.3 Approximation of P_r distribution . . . 74

4.4 Discussion . . . 79

5 Data Driven Models of Synaptic Plasticity 80

5.1 Motivation . . . 80

5.2 Theoretical Foundations 81
5.2.1 Causal States 81
5.2.2 Causal States Transitions 82
5.2.3 Properties of Causal States 82
5.2.4 ϵ-machine Reconstruction 83
5.2.5 Statistical complexity 83
5.3 Causal-State Splitting Reconstruction (CSSR) 84
5.3.1 The Algorithm 84
5.3.2 Parameter selection for CSSR Algorithm 87
5.4 Study of CSSR to Capture the Structure of PSRs 88
5.4.1 Analyzing the map 88
5.4.2 Method 90
5.5 Results 92
5.6 Discussion 98

Bibliography **108**

List of Figures

1.1 A schematic of a biological neuron [12]. 3

1.2 Major elements in chemical synaptic transmission[13]. 5

3.1 Estimated normalized response (P_r) distributions with the control parameter set under stimulation at firing rates A) 0.5, B) 3, C) 8, D) 10, E) 20, and F) 100 Hz. Horizontal axis shows the P_r values and the vertical axis is the relative frequency. 40

3.2 Normalized response (P_r) distribution with the muscarine parameter set under stimulation at firing rates A) 0.5, B) 3, C) 8, D) 10, E) 20, and F) 100 Hz. Horizontal axis shows the P_r values, vertical axis shows the relative frequency. . 41

3.3 Mean of the normalized response P_r distribution plotted against firing rate. . . 43

3.4 Peak of the normalized response P_r distribution plotted against firing rate. . . 43

3.5 Entropy of the normalized response P_r when stimulated with Poisson distributed spike trains of varying mean firing rate for the control and muscarine parameter sets. 45

3.6 Estimated mutual information transmission between normalized postsynaptic response PSR and exponentially distributed preceding $ISIs$ with mean firing rates ranges from 0.1 to 200 Hz. 49

3.7 Information between the postsynaptic response and the preceding sums of interspike intervals in control and muscarine cases. 52

3.8 N-tuple information between inter-spike intervals and postsynaptic response for control and muscarine cases for two frequencies: 5 and 50 Hz. 57

4.1 Histograms of calcium concentration data for mean firing rates 0.5, 3, 8, 10, 20, 100, 5000 and 8000 Hz, together with fitted gamma pdfs. Plots show adherence to a linear relationship between the simulated and theoretical quantiles, confirming our analytic results. 72

4.2 Gamma qq-plot of calcium concentration for 0.5, 3, 8, 10, 20, 100, 5000 and 8000 Hz mean firing rates. 73

4.3 Fixed point values and mean of P_r for the deterministic map as a function of mean firing rates. 74

4.4 PDF of the stochastic fixed point $\overline{PrR}$ for varying interspike intervals of 10, 50, 100 , 120, 330, 2000 in millisecond. Parameters $k_{min} = 0.0013$ and $P_{max} = 0.85$, are from the control set. 77

4.5 Frequency Distribution of P_r for varying mean input ISI A) 10 B) 50 C) 100 D) 120 E) 330 and F) 2000 in milliseconds, when interspike interval T is significantly larger that the calcium decay time τ_{ca} 78

4.6 Quantile plot of two data sets obtained from Pr and $\overline{PR}$. 79

5.1 Fixed point values of normalized postsynaptic response for three synapse models of “depressing”, “mixed”, and “facilitating ” stimulated by Poisson spike trains with mean firing rates ranging from 0.1 to 250. 89

5.2 Fixed point values for release probability P, fraction of readily releasable pool R and normalized postsynaptic response Pr for varying mean firing rates ranges from 0.1 to 100 Hz for **(A)** depressing synapse and from 0.1 to 250 Hz for **(B)** facilitating and **(C)** mixed synapse. 90

5.3 Causal state machines (CSMs) reconstructed and their corresponding relative frequency distributions obtained from depressing FD model. Model is stimulated by Poisson spike trains with mean firing rates **(A)** 0.1, **(B)** 2, **(C)** 5 and **(D)** 100 Hz. The transitions between states are indicated with symbol emitted during the transition (1= large synaptic response, 0 = small synaptic response) and the transition probability. In both **(A)** and **(D)**, CSMs for 0.1 and 100 Hz Poisson spiking process consist of a single state "1" which transitions back to itself, emitting a large response with probabilities 0.9 and 0.06 for low and very high mean firing rates, respectively. In both **(B)** and **(C)**, 2-state CSMs reconstructed for 2 and 5 Hz Poisson spiking process emit large responses with nearly similar probabilities. 93

5.4 Causal state machines (CSMs) reconstructed and their corresponding relative frequency distributions obtained from facilitating FD model driven by Poisson spike train with mean firing rates **(A)** 50, **(B)** 77, **(C)**100, **(D)** 125, **(E)** 200 and **(F)** 250 Hz. State "0" is the baseline state. Similar graph structure is seen for mean firing rates of 50 and 70 Hz. Under mean firing rate of 100 Hz, the graph structure is more complex with more edges, vertices, and one set of parallel edges from state "3" to "6". In nonphysiological range from 125 to 250 Hz, the complexity of graph structure decreases. 95

5.5 Causal state machines (CSMs) reconstructed and their corresponding relative frequency distributions obtained from mixed FD model driven by a Poisson spike train with mean rates **(A)** 5, **(B)** 25, **(C)** 50, **(D)** 125 and **(E)** 250 Hz. In **(A)**, **(C)**, and **(D)**, CSMs for mean firing rates of 5, 50, and 125 Hz consist of two states with similar structure, emitting successive large responses followed by small responses. 3-State CSM for mean firing rate 25 Hz has more complex graph structure. Note that this is inflection point for this synapse model, (see Figure 5.1). 97

5.6 Depressing synapse parameter set. Causal state machines at input frequency of 5 Hz for varying partition threshold. The maximum statistical complexity threshold value, τ, is 0.3 (machine shown in A)), In B) and C) the machine for $\tau + 0.05 = 0.35$ and $\tau - 0.05 = 0.25$, respectively. 105

5.7 Mixed synapse parameter set. Causal state machines at input frequency of 25 Hz for varying partition threshold. The maximum statistical complexity threshold value, τ, is 0.52 (machine shown in A)), In B) and C) the machine for $\tau + 0.05 = 0.57$ and $\tau - 0.05 = 0.47$, respectively. 106

5.8 Facilitating synapse parameter set. Causal state machines at input frequency of 125 Hz for varying partition threshold. The maximum statistical complexity threshold value, τ, is 0.25 (machine shown in A)), In B) and C) the machine for $\tau + 0.05 = 0.3$ and $\tau - 0.05 = 0.2$, respectively. 107

Chapter 1

Introduction

1.1 A Brief History of Neuroscience

The term neuroscience which was introduced in the mid-1960s describes a multidisciplinary science with the goal of analyzing the nervous system and understanding the structure and function of the brain. Neuroscience has attracted researchers from various backgrounds. One probable cause is that the brain can be approached on many different levels, from the purely descriptive to the study of its functional organization. Neuroanatomists study the brain's cellular structure and its circuitry; neurochemists study the chemicals in the brain and its proteins; neurophysiologists are concerned with the study of the brain's properties through recording bioelectrical activity; and theoretical neuroscientists provide a quantitative basis for describing functionality of nervous systems via mathematical models. Neuroscience today incorporates a wide range of research and is one of the most rapidly growing areas of science. Indeed, the brain is sometimes referred to as the final frontier of science[52]. In 1971, 1100 scientists participated at the first Annual Meeting of the Society for Neuroscience (SfN). In 2017, over 30,300 neuroscientists from around the world gathered for the Annual Meeting of

the Society for Neuroscience, the largest annual meeting of scientists in the world at which thousands of abstracts are submitted.

1.2 Biological Background of Basic Neurons

In this section we will review basic functionality of neurons and outline some of their biochemical processes.

1.2.1 Neurons

One of the principle cell type of central nervous system (CNS) in the brain are neurons or nerve cells. The average human brain has 10^{10} to 10^{11} neurons and each neuron can be connected to up to 10^4 other neurons, passing signals to each other through 10^{15} connections. The capabilities of neurons in processing information distinguish them from other cells. A neuron receives information from other neurons in form of electrical signals, processes them, and sends the information to other neurons. A schematic diagram of an ideal neuron is shown in Figure 1.1. There are many different types of neurons, with different sizes and shape. However, there are many common features of neurons which can be understood by describing a generic neuron.

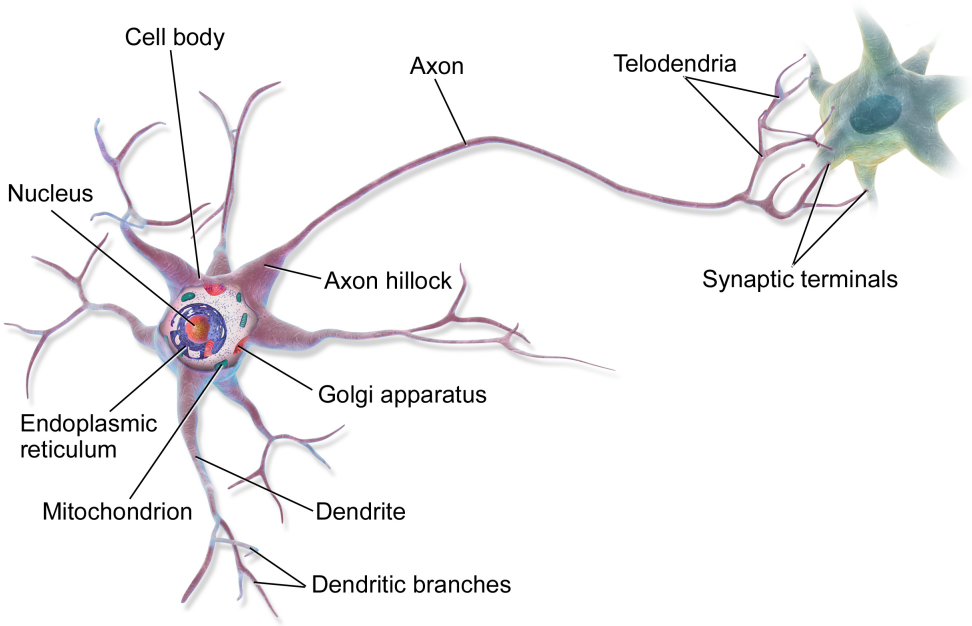


Figure 1.1: A schematic of a biological neuron [12].

1.2.2 Axons

Each neuron has an extension called *axon*, which is a Greek word for axis. It conducts signals away from the cell body to other cells. Some axons can extend as little as one millimeter while others can extend lengths of one meter or more, so they can reach from one region in the brain to almost any other. Axons have a complex biochemical structure to transmit signals over distance based on a phenomenon called an *action potential* or *spike*. Action potentials are fundamental to information processing in neurons.

1.2.3 Biology of Synapses

A synapse, also called neuronal junction, is a contact between neurons, and has a diameter of about $0.5 - 2\mu$m. There are two main types of synapses: electrical and chemical synapses.

Electrical Synapses

Electrical synapses are also known as gap junctions consist of special conducting proteins that allow a direct electromagnetic signal transfer between two neurons. These type of synapses can transmit fast impulses between the connected neurons, although with lower efficiency compare to chemical synapses.

Chemical Synapses

The most common type of connection between neurons in the CNS is the chemical synapse. A presynaptic neuron transmits signals via release of special chemicals called *neurotransmitters*, stored in *synaptic vesicles*, which bind to receptors at the postsynaptic neuron. Many different neurotransmitters have been identified in the nervous system. Common neurotransmitters in the CNS include Dopamine (DA), Glutamate (Glu) or gamma-aminobutyric acid (GABA). The type of neurotransmitters can determine the action on the postsynaptic neuron. This transmitter can be excitatory (common: glutamate) which increases the probability of spiking in the postsynaptic neuron or it can be inhibitory (common: GABA) which decreases the probability of action potential occurring in the postsynaptic cell.

Synaptic Transmission Mechanisms

Upon the arrival of an action potential at presynaptic terminal, depolarization of the axon terminal lead to an influx of calcium through calcium channels. As a result, some of the synaptic vesicles fuse to terminal bouton membrane at special release sites and their neurotransmitters diffuse across the synaptic cleft. Finally these transmitters bind to postsynaptic receptors that generate a postsynaptic response. A schematic diagram of a synapse and the process of action potential arrival at the presynaptic terminal is shown in Figure 1.2.

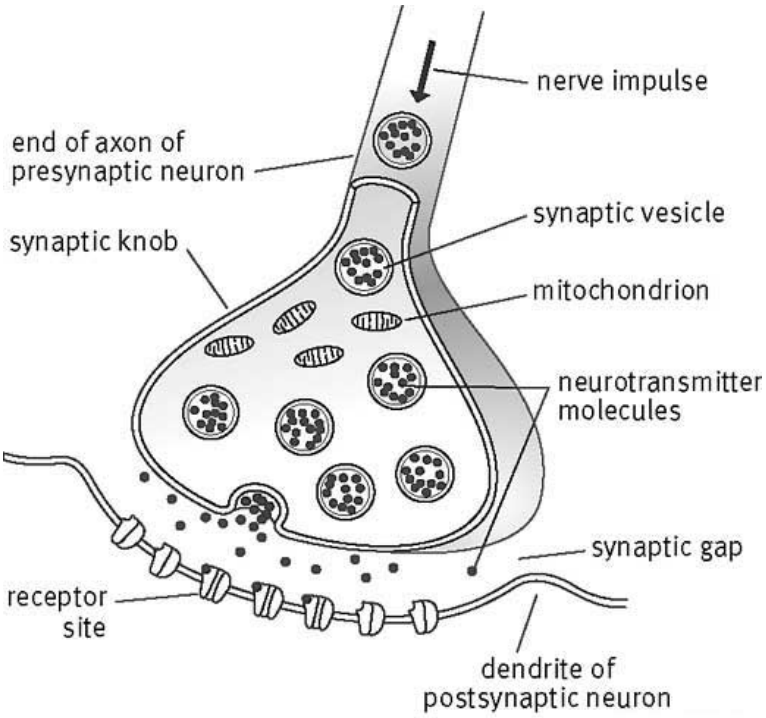


Figure 1.2: Major elements in chemical synaptic transmission[13].

Synaptic Plasticity

One of the most fascinating mechanisms of brain is a biological phenomena called synaptic plasticity. It is the ability of synapses to change their strength and postsynaptic influence in short and long time frames. Synaptic plasticity can modulate the effect of action potentials, either by vesicle release modification of the presynaptic neuron or by changing the sensitivity of postsynaptic neuron. Synaptic plasticity can be divided into two main categories:

- Long-term plasticity (LTP): Long-term changes in synapses that last for hours or longer.
- Short-term plasticity (STP): It allows synapses to change their efficacy on short time scales, milliseconds to minutes. These changes result from vesicle depletion, accumulation of calcium in presynaptic terminal or postsynaptic neurotransmitter receptors

desensitization. There are two principal types of STP, facilitation and depression. Facilitation promotes an increase in the probability of neurotransmitter release(P_r) which then leads to initial growth in postsynaptic current amplitude, whereas depression promotes a decrease in P_r and consequent decline in transmission probability.

Functional roles of STP:

1. Frequency-Filtering: STP conveys frequency filtering properties by making stronger synaptic connections at some firing rates over others. Facilitating synapses are high-pass filters, meaning that they optimally transfer information at high frequencies, depressing synapses are low-pass filters, meaning that they optimally send information at low frequencies, and mixed synapses are band-pass filters.
2. Adaptation and Sensitization : This property allows a neuron to modify its ability to transfer prolonged sensory signals. Adaptation can be caused by synaptic depression which removes the effect of fluctuations in presynaptic excitability while sensitization can be caused by facilitation as it increases the response of a neuron to a given stimulus.

Abbott and Regehr 2004 [1] showed the functional roles of short-term plasticity in GABAergic circuits. GABAergic synapses use GABA the main inhibitory neurotransmitter in the brain. It has an autoreceptor ($GABA_B$) which is located on presynaptic membrane and a postsynaptic receptor ($GABA_A$) which is located at postsynaptic membrane. STP not only depends on the presynaptic activity mentioned earlier on the page, but also depends on feedback activation of presynaptic receptors or postsynaptic processes such as receptor desensitization. For example depression can be partly due to retrograde or feedback inhibitory action of GABA on presynaptic $GABA_B$ receptors by reducing calcium influx to subsequent spikes. This results in a reduction in the future release of this neurotransmitter. Low initial release probability favors facilitation and so the filtering characteristic of the synapse is changed to less depressing and more facilitating. Therefore inhibitory synapses can reduce desensitization. For example, baclofen, which binds to $GABA_B$ receptors, greatly inhibits the initial response to an evoked

stimulus, but later responses are larger and transmission is more reliable in baclofen-treated synapses than non-treated ones during high frequency activation. Therefore, one of advantage of short term depression at inhibitory synapse might be that it helps neurons to perform multiple operations on their inputs. For instance, facilitating inhibitory synapse leads to greater enhancement of release during high frequency bursts.

1.2.4 Neural Coding Theory

How neuronal output is represented relates to the process used by the nervous system to transmit information between cells. Characterizing and analyzing the stochastic components of the neuronal response is a large area of current research. Much of this research focuses on spontaneous action potentials of neurons receiving some stimulus in the form of spike trains. The theory of neural coding models the encoding of sensory messages as action potentials [53]. There are different types of coding depending on how neurons transmit information.

- **Rate Coding**:

  Rate coding or frequency coding was originally formulated by ED Adrian and Y Zotterman in 1926 [4]. It is one way to encode information about the stimulus. Many neuronal models assume that most of the information contained in the neural response can be characterized by its mean firing rate.

- **Temporal Coding**: Unlike rate coding that ignores information from the temporal structure of the spike train, temporal coding assigns importance to temporal precision in the response by focusing on individual spike times and the sequence of interspike intervals [18]. In this case, obtaining the maximal information from response of a neuron depends on how precisely we measure the spike times.

- **Correlation Coding**: This coding scheme assumes that the correlation between spikes may carry additional information, because individual spikes may not encode informa-

tion independently of each other. Therefore in this scheme, we expect that a significant amount of information is carried temporally near spikes. Correlation between spikes makes analysis more complicated. Hence, it is usually assumed that spikes are independent. This assumption can be justified partially by studies showing that the amount of additional information carried temporally near spikes is negligible compared to the information carried by the focal, or principal spike [18].

- **Population Coding**: This coding scheme assumes that information about a stimulus is distributed across the population of responding cells. Therefore, any information about the stimuli or their features are encoded by simultaneous activities of a population of neurons. In studying this type of coding scheme, one must consider not only the firing or temporal patterns of single neurons, but also the relationships of these patterns across responding neurons.

1.3 Tools for Neuroscience: Information Theory and Computational Mechanics

1.3.1 Information Theory

One major goal of this thesis is to understand how the brain stores and conveys information. Indeed, the brain is an input-output system (we react to what we sense) and hence is subject to same laws as other input-output systems. Information theory allows us to answer the following question: " To what extent do neural responses can tell us about the stimuli?". In order to quantify the information transmitted by neurons, we consider the brain as a communication channel since its neurons communicate with each other and transfer information. We use an information theoretic functional developed by Claude Shannon during the 1940s. The theory

provides a mathematical definition of information and describes how much information can be communicated between different parts of a system. His paper published in 1948 [60] and the 1949 book by Shannon and Weaver [61] underpins our understanding of information transfer. Specific details about the elements of information theory and how they apply to measuring information at the synapse level are presented in the next Chapters.

1.3.2 Computational Mechanics

Computational mechanics, (CM) introduced by James Crutchfield in 1989 [17], is an approach to address the issue of pattern, structure and organization observed in data. It illustrates how to formulate a model of the hidden process that generates observed behavior and then extrapolate beyond the original observational data to make predictions of future behavior [57]. In a nutshell, the goal of CM is to identify patterns which are most informative about a hidden process without assuming 'a priori' patterns for the observed data, proceeding from pattern analysis to pattern discovery. The CM procedure is to discretize the empirical data into finite alphabet and aim for "causal states". Two series of past data, two histories, with the same distribution of future data leave one in the same causal state. In other words, if there is no statistically significant difference between these two histories in the future, they are in the same causal state. This procedure identifies the causal states, the structure of the connections and succession in causal states. The automata created is called an ϵ-*machine* and the procedure ϵ-*machine reconstruction*. The name might be somewhat strange, but I have not heard better one. We describe CM procedure in detailed in the chapter 5 and explore the potential of the approach in the analysis of the neuronal data.

Chapter 2

Information Theory

2.1 Information theory in neuroscience

After the publications of Shannon's paper "A mathematical theory of communication", several researchers begin applying information theory in neuroscience. MacKay and McCulloch in 1952 [44] investigated the capacity of a neural cell for transmitting signals using the concept of information. This work suggested the later work on understanding how much information flows through nervous system. These concepts are referred to as "Neural Information Flow" and it has been both highly versial and influential in the neuroscience community. Stein in 1967 [63] examined the information capacity of nerve cells using a frequency code and clarified the discrepancies between timing versus frequency coding, still one of the major debates among neuroscientists. In [8], information transmission is studied through a master equation based on stochastic model of presynaptic release of vesicles combined with a low dimensional model of membrane charging at the post-synaptic side. In 2002, Fuhrmann and his collaborators [23] characterize synaptic transmission in the neocortex and quantify the amount of information conveyed by a single response to a specific sequence of spike stimulation, as it is effected

by short term synaptic plasticity. They reported that for a given dynamic synapse there is an optimal frequency of input stimulation for which the information transfer is maximal. A mathematical model of the calyx of Held was used in [76] to study synaptic depression due to repeated stimulation in vitro. They compute the information contained in the postsynaptic current amplitude about preceding interspike intervals using information theoretic measures. Part of this thesis is advances by the work of Fuhrmann et al. [23] and Yang et al. [76].

2.2 Information-Theoretic Functionals

Recall from Chapter 1 that a milestone of science is the theory of information introduced by Claude Shannon [60]. Shannon developed a mathematical theory that quantifies uncertainty over noisy communication channels. Probability theory is a mathematical framework that provides a means of quantifying uncertainty and axioms for deriving uncertainty statements. In this section we describe information theoretic functionals, as well as the definitions and notations of probabilities, in particular for those readers with limited exposure to probability theory.

2.2.1 Probability Distributions and Densities

Random Variables

A random variable is a function that maps the outcomes of a probabilistic experiment to the real numbers. We typically denote the random variable with upper case letter and the values it can take (realizations) with lower case letters. For example x_1 and x_2 are two realizations of the random variable X. Random variables can be discrete or continuous. A discrete random variable can take on countably infinite number of values, while a continuous random variable may take an uncountably many values.

Probability Distributions

A probability distribution describes how likely a random variable or a vector of random variables (otherwise known as a multivariate r.v.) is to take on each of its possible realizations. We describe probability distributions for both discrete and continuous structure of random variables.

Discrete Variables and Probability Mass Functions

A probability distribution over discrete random variables is described using a *probability mass function* (PMF). It is typically denoted by P. A probability mass function may be defined for multivariate random variables. Such a probability distribution is known as *joint probability distribution* and it is denoted for example, by $P(X = x, Y = y)$ or $P(x, y)$. Probability mass functions have the following properties:

- The domain of P is the set of all possible values of X.
- $\forall x \in X,\ 0 \leq P(X = x) \leq 1$.
- $\sum_{x \in X} P(X = x) = 1$.

Continuous Variables and Probability Density Functions

A probability distribution over continuous random variable is described using a *probability density function* (PDF) and it is denoted by $f_X(x)$. To be a probability density function, a function must satisfy the following properties:

- The domain of f is the set of all possible values of X.

- $\forall x \in X,\ f_X(x) \geq 0$. Note that the probability density function can take on values greater than one.

- $\int f_X(x)dx = 1$.

  If $\mathbf{X}$ is a multivariate random variable, then $\int \int \cdots \int f_X(x_1, \cdots, x_p)dx_1 \cdots dx_p = 1$

- $P(X \in [x, x+dx]) = f_X(x)dx$, which is the probability of landing inside an infinitesimal region with volume dx.

Marginal Probability

The probability of the occurrence of the single event is known as the *marginal probability distribution.*

For discrete case marginal probability is calculated using sum over the joint probability:

$$\forall x \in D(X),\ P(X = x) = \sum_y P(X = x, Y = y).$$

For continuous variables, we use integration instead of summation

$$f_X(x) = \int_{D(X)} f(x, y)dy.$$

Conditional Probability

When we are interested in probability of an event, given that another event has happened, we use the *conditional probability distribution.* For discrete case it is denoted by $P(X = x|Y = y)$ and can be computed with the following formula

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

For continuous case it is denoted by $f(x|y)$ and can be computed with the following formula

$$f(x|y) = \frac{f(x,y)}{f(y)}.$$

2.2.2 Entropy

A fundamental quantity in information theory is called *Entropy*. It is the amount of uncertainty associated with a random variable. The entropy of a discrete random variable X with PMF P(x) is

$$H(X) = -\sum_x P(x) \log_2 P(x), \tag{2.2.1}$$

where, by convention, base 2 logarithms are used, however the choice of base is somewhat arbitrary, as it is relatively easy to convert from one form to other. To indicate that the base 2 logarithm is being used, information is reported in units of "bits". $H(X)$ can be defined as the expected uncertainty in a random variable X

$$H(X) = -\sum_x P(x) \log_2 P(x) = -\mathbb{E}[\log_2 P(X)]. \tag{2.2.2}$$

2.2.3 Joint Entropy

Consider two random variables X and Y jointly distributed according to the PMF $P(x,y)$. Their joint entropy [15] is

$$H(X,Y) = -\sum_x \sum_y P(x,y) \log_2 P(x,y). \tag{2.2.3}$$

The joint entropy measures the amount of uncertainty in the two random variables X and Y taken together. We can generalize joint entropy to an arbitrary number of random variables

and obtain the following

$$H(X_1, X_2, \cdots, X_n) = -\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} P(x_1, x_2, \cdots, x_n)\log_2 P(x_1, x_2, \cdots, x_n). \tag{2.2.4}$$

2.2.4 Conditional Entropy

Given two random variables X and Y, their conditional entropy [15] is defined as

$$H(X|Y) = -\sum_{x}\sum_{y} P(x, y)\log_2 P(x|y). \tag{2.2.5}$$

The conditional entropy is a measure of how much uncertainty remains about the random variable X, when we know the value of Y.

From this equation we can drive the following identity

$$\begin{aligned} H(X|Y) &= -\sum_{x}\sum_{y} P(x, y)\log_2 \frac{P(x, y)}{P(y)} \qquad (2.2.6)\\ &= -\sum_{x}\sum_{y} P(x, y)\left(\log_2 P(x, y) - \log_2 P(y)\right)\\ &= -\sum_{x}\sum_{y} P(x, y)\log_2 P(x, y) - \left(-\sum_{x}\sum_{y} P(x, y)\log_2 P(y)\right)\\ &= -\sum_{x}\sum_{y} P(x, y)\log_2 P(x, y) - \left(-\sum_{y} P(y)\log_2 P(y)\right)\\ &= H(X, Y) - H(Y) \qquad (2.2.7) \end{aligned}$$

Some of the properties of the entropic quantities defined above are as follow:

- Non negativity: $H(X) \geq 0$. This quantity is zero if and only if X is deterministic.
- Monotonicity: Conditioning always reduces entropy:

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

- Entropy is unrelated to the temporal structure of the data. In other words, entropy does not change by randomly shuffle around the time points. For example, two signals such as random noise and sine wave might be very different and it can be due to their temporal structure and not their distributions of data values and hence these two signals can have similar entropy based on the " timeless" distribution and not on the temporal structure.

2.2.5 Mutual Information

Mutual information between two discrete random variables X and Y is given by

$$I(X;Y) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{2.2.8}$$

$$= H(X) - H(X|Y)$$

$$= H(X) + H(Y) - H(X,Y). \tag{2.2.9}$$

Note that for any random variables, X and Y, $I(X;Y) \geq 0$, with equality if and only if X and Y are independent [15].

It can be seen from eqn. (2.2.9) that the mutual information is the difference between uncertainties about X before and after observing Y so it can be defines as the amount of reduction in uncertainty in one variable after having knowledge of another variable.

Some of the properties of mutual information are as follows:

- Mutual information is symmetric, i.e. $I(X;Y) = I(Y;X)$.
- For a given $H(X)$, $I(X;Y)$ is maximum when $H(X|Y) = 0$, i.e. X is completely defined by Y. In this case $I(X;Y) = H(X)$. Therefore, we can interpret entropy as the maximum amount of information that can be gained about a random variable and it is sometimes called *self-information* of X.

- $I(X:Y) = 0$ If and only if X and Y are two independent random variables.
- While the Pearson correlation coefficient does not permit the detection of non-linear relationships, mutual information can be used to measure non-linear associations.

In this thesis, mutual information is employed to estimate the amount of information a neuronal postsynaptic response carries about the preceding presynaptic interspike intervals in a model of activity-dependent GABAergic synapses.

2.2.6 Differential Entropy

The concept of entropy can be generalized to continuous random variables. However, it should be noted that the discrete Shannon entropy as defined in eqn. (2.2.1) is not an approximation of the analogous continuous entropy [15]. In other words, the continuous Shannon entropy cannot be derived by discretizing the continuous random variable and letting the number of intervals tend to infinity and passing to the limit. To perceive how this issue emerges, assume X is a continuous random variable with probability density function $f_X(x)$. Suppose that one could discretize this variable with precision Δx across the whole domain of X. It is clear that the probability of observing values of X in an interval of width Δx centered around x_i is $P_i = f_X(x_i)\Delta x$, where $\sum_i P_i = 1$. The entropy of discretized version of X, which we denote by X^Δ, is given by

$$\begin{aligned} H(X^\Delta) &= -\sum_i f(x_i)\Delta x \log_2 \left(f(x_i)\Delta x\right) \\ &= -\sum_i f(x_i)\Delta x \log_2 f(x_i) - \log_2 \Delta x \sum_i f(x_i)\Delta x \\ &= -\sum_i f(x_i)\Delta x \log_2 f(x_i) - \log_2 \Delta x \end{aligned}$$

As long as $f(x)\log_2 f(x)$ is Riemann integrable, we know that as Δx approaches zero the first term on the right hand side becomes the integral, that is,

$$-\sum_i f(x_i)\Delta x \log_2 f(x_i) \xrightarrow{\Delta x \to 0} -\int f(x)\log_2 f(x)dx$$

Thus we see that

$$H(X^\Delta) = -\int f(x)\log_2 f(x)dx - \log_2 \Delta x \tag{2.2.10}$$

As $\Delta x \to 0$ the entropy diverges

$$H(X) = -\int f(x)\log_2 f(x)dx + \infty,$$

which shows that knowing a continuous quantity requires an infinite amount of information. Therefore, it is not feasible to obtain the entropy of continuous random variable based on Shannon entropy. However, it is practical to compute the difference between entropies of continuous random variables by taking the limit $\Delta x \to 0$, assuming same precision Δx for both variables, because the term $\log_2 \Delta x$ will cancel out. Therefore, the difference between entropies of two random variables X and Y is given by

$$H(X^\Delta) - H(Y^\Delta) = -\sum_i f(x_i)\Delta x \log_2 f(x_i) + \sum_i f(y_i)\Delta x \log_2 f(y_i)$$

A measure of entropy called *differential entropy* of continuous random variable X with PDF $f_X(x)$ ignores the divergent term; it is defined as

$$h(X) = -\int f_X(x)\log_2 f_X(x)dx. \tag{2.2.11}$$

For more details regarding the limiting process, see [15].

Recall from Chapter 3.2.1 that because probability mass function $0 \leq P(X = x) \leq 1$

$\forall x \in D(X)$, then $H(X)$ is always nonnegative. However, for a continuous random there is no constraint on the probability density function $f_X(x)$ and it can take on values greater than 1. This means it is possible that $h(X)$ be negative which is counterintuitive.

Mutual information between two continuous random variables with same precision Δ is the difference between differential entropies and it is defined as [15]

$$\begin{aligned} I(X;Y) &= \int_x \int_y f(x,y) \log_2 \frac{f(x,y)}{f(x)f(y)} dxdy \\ &= h(Y) - h(Y|X) \\ &= \left(H\left(Y^\Delta\right) + \log_2 \Delta\right) - \left(H\left(Y^\Delta|X^\Delta\right) + \log_2 \Delta\right) \\ &= H\left(Y^\Delta\right) - H\left(Y^\Delta|X^\Delta\right) \\ &= I(X^\Delta;Y^\Delta) \end{aligned} \tag{2.2.12}$$

Since $I(X^\Delta;Y^\Delta) = I(X;Y)$, properties of discrete mutual information such as being nonnegative and symmetric extend to continuous mutual information.

2.3 Estimation of Mutual Information

Quantifying the amount of information neuronal activity carries helps us to better understand the statistical features of spike trains such as timing of spikes and interspike interval patterns. However, estimation of information from empirical data can be problematic and in many instances, there are not a known families of distributions that can describe the experimental data. In such situations, a parametric approach to estimation of mutual information is not possible and hence non-parametric approaches are recommended. Here, we consider two non-parametric classes of information estimators.

2.3.1 Partition Based Estimators

One of the traditional methods to estimating probability densities is the histogram. The idea is to count the number of data points that fall into each interval of a certain partition of the domain of the random variable. Roughly speaking this procedure corresponds to the maximum likelihood estimator of probability densities. The corresponding entropy estimate is written

$$\hat{H}_{MLE} = -\sum_{x} \hat{P}(x) \log_2 \hat{P}(x),$$

where the unknown probability distribution $P(x)$ is replaced by empirical probabilities $\hat{P}(x)$. This estimator is often called naive or maximum likelihood estimator (MLE) after the fact that $\hat{P}$ is the maximum likelihood estimator of P in the case of a discrete random variable X. Estimation of information theoretic functionals from histogram-based probability density models depends on the choice of the number of intervals. While an optimal choice requires knowledge about the underlying probability density function f, this knowledge is usually rare. Instead, one suggestion is to choose the widths of the histogram intervals sufficiently large to capture the major features in the data and ignore fine details due to random sampling variations [35]. Intervals that are too large fail to describe a sharply peaked probability density function and will underestimate the information functionals. Conversely, intervals that are too small lead to biases associated with imprecision and will overestimate the functionals. Several guidelines exist for selecting the number of intervals. *Sturge's rule* [67] is merely based on the number of sampled points n and it is given by

$$nbins = 1 + \log_2 n. \tag{2.3.1}$$

For moderate number of data points (less than 200) *Sturge's rule* produces reasonable histograms when data are normally distributed and symmetrical, so it maps the data into discrete, symmetric, binomial classes. For an extremely large number of data points or a severely skewed data, this method is not recommended. This is partly because the method only con-

siders the number of data points and not the range of the distribution, and this can lead to oversmoothed histograms and hence underestimates the appropriate number of bins [56]. Scott [55] formulated a data-based choice for the number of bins which asymptotically minimizes the integrated mean square error of a histogram estimate, $\hat{f}(x)$, of the true density value, $f(x)$,

$$IMSE = \int \mathbb{E}\left(\hat{f}(x) - f(x)\right)^2 dx,$$

and it is given as

$$nbins = \frac{R \times n^{1/3}}{3.49 \times s}, \tag{2.3.2}$$

where R is the range of sampled data and s is an estimated standard deviation. Scott considers Tukey's [69] suggestion to use the Gaussian density as a reference standard cautiously, but frequently and assumed that data are normally distributed. Freedman and Diaconis [22] reported a simple robust rule for choosing cell width. The *Freedman and Diaconis Rule* (*FD Rule*) is to choose the bin width as twice the interquartile range of the data divided by the cube root of the sample size

$$nbins = \frac{R \times n^{1/3}}{2 \times IQR},$$

where $IQR = Q_3 - Q_1$ is the interquartile range which is the difference between 25th and 75th percentiles of sample points. Unlike *Scott's Rule*, the *Freedman and Diaconis Rule* does not make any assumption about the underlying density function and hence it can be applied to any data distributions. In the case of normal distribution both rules provide similar results. The methods of selecting optimal number of bins outlined above are considered for one variable so they can be applied to marginal entropy estimations. However, in estimating mutual information or multivariate entropies based on histograms, care should be taken as the optimal number of bins can be different for each variable. To address this issue, it is recommended [10] to find the optimal number of bins using the selected method of interest for each variable,

compute the ceiling of the average of the optimal number of bins, and apply this to all variables.

Bias

Information estimates based on binning are sometimes contaminated by bias. This bias is caused by both an inadequate representation of the probability density function using a histogram, and insufficient sample size. In fact, from Jensen's Inequality it can be proved that the entropy estimates obtained from the maximum likelihood estimator (MLE) of the probability density function is negatively biased unless P is trivial.

$$\mathbb{E}_p\left(\hat{H}_{MLE}\right) \leq H(P).$$

In other words, we have equality in the above expression only when $H(P) = 0$; in words, the bias of the MLE for entropy is negative unless the underlying distribution P is supported on a single point.

Theorem. *The Maximum likelihood estimator for the entropy ($\hat{H}_{MLE}$) is negatively biased Everywhere [51].*

Proof. In the context of the probability theory, if X is an integrable real-valued random variable and ϕ is a concave function, then

$$\mathbb{E}\left(\phi(X)\right) \leq \phi\left(\mathbb{E}(X)\right).$$

Since Entropy is a concave function [15], then we have

$$\mathbb{E}\left(H(\hat{P})\right) \leq H\left(\mathbb{E}(\hat{P})\right), \tag{2.3.3}$$

where $\hat{P} = \frac{n_x}{n}$ is a maximum likelihood estimator of P, n_x is the number of success in Bernoulli

trials. Since

$$\mathbb{E}(\hat{P}) = \mathbb{E}(\frac{n_x}{n}) = \frac{1}{n}\mathbb{E}(n_x) = \frac{1}{n}(nP) = P,$$

therefore $\hat{P}$ is an unbiased estimator of P and we can rewrite (2.3.3) as

$$\mathbb{E}\left(H(\hat{P})\right) \leq H(P)$$
$$\mathbb{E}(H(\hat{P})) - H(P) \leq 0$$
$$Bias(\hat{H}) \leq 0.$$

Therefore, ML estimator (also called "plug-in" - by Antos & Kontoyiannis, 2001 [5])will always be negatively biased unless the underlying distribution P is supported on a single point in which $H(P) = 0$. □

It was shown in [30, 48, 25] that the bias of entropy and mutual information associated with sample sizes can be estimated and the expectation of $\hat{H}_{MLE}$ is given by

$$\mathbb{E}\left(\hat{H}_{MLE}\right) = H - \frac{M-1}{2n\ln 2},$$

where M is the number of bins. The mutual information is a sum of entropies, so we extend expression 2.3.4 to estimate the bias of mutual information [30] given by

$$\mathbb{E}\left(\hat{I}(X;Y)\right) = I(X;Y) + \Delta I(X;Y),$$

with $\Delta I(X;Y) = \frac{M_{xy}-M_x-M_y+1}{2n\ln 2}$. Here M_x, M_y, M_{xy} denote the number of bins for marginal and joint variables X and Y. However, as it was mentioned earlier, it is recommended to use same number of bins for different variables as choice of bin width influence the entropy estimation. In this case, if $M = M_x = M_y$ and $M_{xy} = M \times M$, then we have

$$\mathbb{E}\left(\hat{I}(X;Y)\right) = I(X;Y) + \frac{(M-1)^2}{2n\ln 2}, \tag{2.3.4}$$

Note that while the above bias corrections can be helpful, no bias correction is effective when the amount of data is very limited. In fact, partition-based estimators suffer from the "curse of dimensionality", a problem related to the sparsity of the available data when the number of random variables is large [74]. This problem exists due to limited amounts of data in the analysis of physiological systems. In other words, the number of bins (M) in a regular partition grows exponentially (M^{dim}) with the number of dimensions of the data and can exceed the number of observations n. The result in sparse data with many empty bins leading to large biases for information functionals. It is possible to consider adaptive partitions where cells vary to accommodate the distribution of observations. Optimized estimators use adaptive bin sizes having equal numbers of observations $n(i, j)$ for all pairs. However, these estimators may still have systematic errors resulting from approximating $I(X, Y)$ by $I_{binned}(X, Y)$, and the approximation of probabilities by relative frequencies.

2.3.2 Metric Based Estimators

It is ideal to find an asymptotically unbiased estimator of entropies of continuous probability distribution from a sample of observations on a Euclidean vector space that can avoid the difficulties associated with binning. Kozachenko and Leonenko [36] developed an entropy estimator, referred as to the "KL estimator" based on a nearest neighbor search. They show that for a finite sample drawn from a continuous probability distribution in a Euclidean space, k-nearest neighbor (k-NN) distances (i.e. the distance from a point to its kth nearest neighbor amongst the sample points, in some metric on the space) provide an asymptotically unbiased and consistent estimator of entropy. This class of estimators is based on the notion that the larger the distance between one point to its nearest neighbor, the smaller the local density is around that point.

The KL estimator implies that to estimate information theoretic functionals it is not necessary to explicitly estimate the full probability density function from the sample observations. A simple example is when estimating the mean of n sample points without estimating its

probability distribution. Instead we easily apply the formula $\frac{1}{n}\sum x_i$ which provides direct estimation of the mean from samples without estimating the underlying distribution. In what follows, we refer to Kozachenko and Leonenko [36] estimator as "KL entropy".

Kozachenko-Leonenko (KL) entropy estimates of one dimensional distributions

Let X be a continuous random variable with values defined in a metric space, i.e., there is a distance function $\|x - x'\|$ between any two realizations of X. Let $f(x)$ be a probability density function on a real line X. The goal is to estimate differential entropy defined by

$$h(X) = -\int_{x \in D(X)} f(x) \log_2 \left(f(x)\right) dx \tag{2.3.5}$$

from a sample of realizations $x_1, \cdots, x_n$ drawn from $f(x)$.
We seek an estimate for differential entropy that depends continuously on observations. We exploit the continuous nature of $f(x)$, but we keep the estimation procedure local so that the sensitivity to the shape of $f(x)$ is preserved.

The first step is to notice that a differential entropy term (2.3.5) can be rewritten as

$$h(X) = -\mathbb{E}\left(\log_2 f(X)\right).$$

Hence, we can approximate differential entropy h by the sample mean of $\log_2 f(x)$ evaluated at the points $x = x_i, i = 1, \cdots, n$ without explicitly estimating its PDF. The approximation is given by

$$h(X) \approx \frac{-1}{n} \sum_{i=1}^{n} \log_2 f(x_i).$$

Therefore, if we can find an unbiased estimator for $\log_2 f(x)$, say $\widehat{\log_2 f(x)}$, then we have an

unbiased estimator for differential entropy $h(X)$. Since we have

$$\begin{aligned}\hat{h}(X) &= -\frac{1}{n}\sum_{i=1}^{n}\widehat{\log_2 f(x_i)} \\ \mathbb{E}\left(\hat{h}(X)\right) &= -\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\widehat{\log_2 f(X_i)}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log_2 f(x_i) \\ &= h(X).\end{aligned}$$

In order to obtain the unbiased estimator $\widehat{\log_2 f(X_i)}$, we assume $Q_k(\epsilon)$ to be the probability distribution for the distance between x_i and its kth nearest neighbor. The distance to the kth nearest neighbor (kNN) can be seen to be related to a local density estimate since the larger the distance is to kNN, the smaller the local density. Therefore, $Q_k(\epsilon)\,d\epsilon$ can be treated as if it were the probability that there is one point within the infinitesimal distance $r \in [\epsilon, \epsilon + d\epsilon]$ from x_i, that is $k-1$ other points at smaller distances, and that $n-k-1$ points have larger distances from x_k.

Let q_i be the mass of the ϵ ball centered at x_i so we have

$$q_i(\epsilon) = \int_{\|\eta - x_i\| < \epsilon} f(\eta) d\eta.$$

Using trinomial formula we obtain

$$Q_k(\epsilon)\,d\epsilon = \frac{(n-1)!}{1!(k-1)!(n-k-1)!} \times \frac{dq_i(\epsilon)}{d\epsilon} \times d\epsilon \times q_i^{k-1} \times (1-q_i)^{n-k-1},$$

or

$$Q_k(\epsilon) = k \times \binom{n-1}{k} \times \frac{dq_i(\epsilon)}{d\epsilon} \times q_i^{k-1} \times (1-q_i)^{n-k-1}.$$

Note that $Q_k(\epsilon)$ is a probability density function and it can be easily checked that $\int_0^\infty Q_k(\epsilon)\,d\epsilon =$

1. We can also compute

$$\begin{aligned}\mathbb{E}\left(\log_2 q_i\left(\epsilon\right)\right) &= \int_0^\infty \left(\log_2 q_i\left(\epsilon\right)\right) Q_k\left(\epsilon\right) d\epsilon \\ &= \frac{k\binom{n-1}{k}\Gamma(k)\Gamma(n-k)\left(\psi^0(k)-\psi^0(n)\right)}{\Gamma(n)} \\ &= \psi(k)-\psi(n),\end{aligned}$$

where $\psi(x)$ is digamma function which is defined as the logarithmic derivative of the gamma function, $\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}$. Assuming a uniform local density in a small environment around each sample point, we can approximate $f(x_i)$ as follows:

$$f(x_i) \approx \frac{q\left(\epsilon_i\right)}{V},$$

where V is the volume surrounding each sample point x_i and is formulated as

$$V = c_d \times \epsilon_k^d(x_i),$$

where $c_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$ is the volume of unit sphere in d dimension and $\epsilon_k(x_i)$ is the distance between the estimation point x_i and its kth closest neighbor. Therefore, we have

$$q\left(\epsilon_i\right) \approx c_d \times \epsilon_k^d \times f(x_i)$$

and

$$\log_2 f(x_i) \approx \log_2 q(\epsilon_i) - \log_2 c_d - d\log_2 \epsilon_k.$$

So

$$-\mathbb{E}\left(\widehat{\log_2 f(x_i)}\right) = -\mathbb{E}\left(\log_2 q(\epsilon_i)\right) + \log_2 c_d + d\mathbb{E}\left(\log_2 \epsilon_k(i)\right)$$
$$h(X) \approx -\psi(k) + \psi(n) + \log_2 c_d + \frac{d}{n}\sum_{i=1}^{n} \log_2 \epsilon_k(i)$$

Therefore,

$$\hat{h}(X) = -\psi(k) + \psi(n) + \log_2 c_d + \frac{d}{n}\sum_{i=1}^{n} \log_2 \epsilon(i). \tag{2.3.6}$$

Eqn. (2.3.6) is unbiased if density $f(x_i)$ is uniform. It is shown in [36, 39] that the bias of the underlying estimates is caused by non uniformity of the underlying density.

Similarly, let $Z = (X, Y)$ be a random variable and let $\|.\|$ be maximum norm defined as $\|z - z'\| = max\{\|x - x'\|, \|y - y'\|\}$ while any norm can be used for X and Y, d and c_d in eqn. (2.3.6) are replaced by $d_x + d_y$, and $c_{d_x} c_{d_y}$, respectively. Therefore,

$$\hat{h}\left(X, Y\right) = -\psi(k) + \psi(n) + \log_2 c_{d_x} c_{d_y} + \frac{d_x + d_y}{n}\sum_{i=1}^{n} \log_2 \epsilon_k(i). \tag{2.3.7}$$

From eqn. (2.3.6) and (2.3.7) we can estimate mutual information in terms of entropies for the fixed k.

The above estimator is asymptotically unbiased and consistent. The consistency of this estimator has been proven for $k = 1$ by the original authors [37] and for general k by [62]. The bias is minimal compare to partition-based and plug-in estimators [47]. Victor [71] showed that by using the KNN approach, data efficiency can be as much as 1000 times greater than that of histogram strategies for electrophysiological data sets. Also, the KL estimator applies an adaptive resolution since the distance scale depends on the underlying density function. In addition, the search for the closest neighbor is a classical problem that has received a lot of attention and for which several algorithms exist that make the procedure computationally more efficient [29, 77]. However, the KL estimator will suffer from the curse of dimensionality

if the small sample size in high dimensional space. Also, it is important to note that the probability densities involved in computing mutual information from individual terms are of different dimensionality. Therefore, for fixed k, different distance scales are used for spaces of different dimensions. For instance, the distance to the kth neighbor in the joint space tends to be larger than the distance to the neighbors in the marginal spaces. As a result, the size of the ϵ-ball within which we assume that the density of the sample distribution is constant depends directly on the dimensionality of the samples. Therefore, the biases of the estimated entropies which depends on the validity of this assumption would be non-zero as they would be different in $\hat{h}(X)$, $\hat{h}(Y)$, and $\hat{h}(X,Y)$. This might lead to biased estimation. To avoid this problem, Kraskov and his collaborators [39, 38] provided a methodology to adapt the KL estimator to estimate mutual information. Their estimator is referred to as the KSG estimator. The KSG estimator estimates mutual information via KL estimator for entropies with a choice of nearest neighbor parameter k. In particular, in this algorithm k is varied for each sample point so that the radii of the corresponding ϵ-ball is approximately the same for the joint and the marginal spaces. Therefore, we use a fixed k only in the higher dimensional space and project the distance scale set by this k into the lower dimensional spaces. Then an estimate for mutual information is obtained by counting the number of sample points ($n_x(i)$ or $n_y(i)$) that fall within a set distance for each point in the marginal space. Hence the KSG entropy estimator is

$$\hat{h}(X) = -\frac{1}{n}\sum_{i=1}^{n}\psi[n_x(i)+1] + \psi(n) + \log_2 c_{d_x} + \frac{d_x}{N}\sum_{i=1}^{n}\log_2 \epsilon(i).$$

Note that the systematic biases in $\hat{h}(X)$, $\hat{h}(Y)$, and $\hat{h}(X,Y)$ will not cancel exactly, however the probability that they will approximately cancel are greater with the KSG procedure than had we used different length scales in the three estimates. The real proof comes of course from detailed numerical tests [39].

The estimate of mutual information between two random variables obtained from the KSG

algorithm is then [39]

$$\hat{I}(X;Y) = \psi(k) - < \psi(n_x + 1) + \psi(n_y + 1) > + \psi(n),$$

where $< \cdots >= \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}}\,(\cdots(i))$ is the averages of n_x and n_y over all samples.

The KSG estimator is directly extendible to multi-dimensional information [39] and can be written as

$$\hat{I}(X_1, X_2, \cdots, X_m) = \psi(k) + (m-1)\psi(n) - < \psi(n_{x_1}) + \psi(n_{x_2}) + \cdots + \psi(n_{x_m}) >,$$

Note that both of the KL and KSG algorithms for information estimator are implemented in the Java Information Dynamics Toolkit (JIDT) [42]. The code was written in Java but it can be called directly from MATLAB, R, and Python.

In brief, KSG estimator is more data efficient and accurate compare to other methods such as histograms and so it allows us to analyze limited and possibly noisy experimental data sets. Also, the KSG algorithm is more resistant to the biases associated with high dimensional variables better than histogram techniques. Therefore, with limited data originating from an unknown distribution is common in neuroscience, the KSG improves the applicability of information theoretic functionals.

In the next Chapter, we focus on the use of partition-based and KSG estimators using a electrophysiological data set.

Chapter 3

Information Processing in Hippocampal Interneuron Synapses

Changes in the strength of synaptic connections are attributed either to the depletion of readily releasable vesicles, leading to synaptic depression, or from an increased probability of releasing neurotransmitter, leading to synaptic facilitation. These changes in the temporal structure of neuronal activity consequently effect the magnitude of postsynaptic response. Therefore, we expect that postsynaptic response carries some information about the temporal pattern of presynaptic activity.

The critical role of the hippocampus in support of memory and learning processes cannot be ignored. Recent studies showed that temporal coding is a strong aspect of hippocampal firing patterns [32]. The goal of this chapter is to quantify the amount of information contained in the postsynaptic response induced by a Poisson spike train about the preceding temporal activity such as interspike intervals in pairs of synaptically connected neurons.

3.1 Motivation

Several studies have investigated the amount of information transfer at synaptic level. For instance, in [76] the mathematical model of calyx of Held was used to compute information transmission at the glutamatergic synapse. They determined that the amplitude of the response decreases as the mean firing rate (the number of spike per unit of time) increases. Markram and his collaborators in [23] consider synaptic transmission in the neocortex, and estimate the amount of information carried by a single response to a sequence of interspike intervals. They showed that for any given dynamic synapse, there is an optimal frequency of input stimulation for which the information content is maximal. We begin this chapter by introducing a model of facilitation and depression (FD) [66] that was fit using experimental data recorded from a hippocampal inhibitory GABAergic interneuron pyramidal cell connections. For fixed frequency input spikes at frequencies in the range of theta ($\sim$ 4-12 Hz) and gamma ($\sim$ 20-100 Hz) oscillations. These results apply to the micro-circuits in the hippocampus that are responsible for the interaction of theta and gamma frequencies associated with learning and memory. A control state was compared to one where a pharmaceutical muscarinic neuromodulator was applied. Next, we apply information theoretic functionals to the facilitating and depression model to investigate the information processing properties of this synapse under control and neuromodulation conditions. We examine frequencies that range from near zero to 100 Hz because gamma and theta brain rhythms and these rhythms carry distinct functions and allow the brain to take in and to learn new information [11]. We then apply techniques that measure the dependence of the response on the exact history of presynaptic activity. This effort reveals some unexpected distinctions between control and muscarine-added cases. Finally, we end this chapter with a conclusion and brief discussion of techniques. It should be noted that both [23, 76] have directly inspired the work presented in this chapter.

3.2 FD model

Whole-cell recordings from synaptically connected pairs of neurons in mouse hippocampal slices from PV-GFP mice were performed [40]. The presynaptic neuron was a PV basket cell (BC) and the postsynaptic neuron was a CA1 pyramidal cell. In this experiment they used 1-2 ms duration supra-threshold current steps in order to evoke action potentials in the PV-BC from a resting potential of -60 mV. Also, trains of 25 action potentials are evoked at 5, 50, and 100 Hz from the presynaptic basket cell. The outcome in the postsynaptic neuron was the activation of $GABA_A$-mediated inhibitory postsynaptic currents (IPSCs). Upon repeated stimulation, the amplitude of IPSC decreases to a steady-state level. It should be noted that muscarine binds and activates presynaptic acetylcholine muscarinic receptors (mACHRs) which lead to inhibition of presynaptic calcium channels. This decreases the amount of calcium that floods the terminal upon arrival of the action potential and causes a reduction in the response overall and the amount of depression in the train. More details of this experiment can be found in [40].

Stone and colleagues in [66] parametrize a model of presynaptic plasticity with the experimental data obtained from [40]. In this model the release probability P_{rel} is the fraction of a pool of synapses that release a vesicle upon arrival of a spike at the synaptic terminal. Following the release of vesicles upon stimulation, a portion of synaptic sites cannot release vesicles for a certain period of time. Repeated simulation causes a depletion of the vesicle pool which leads to depression.

The peak of measured IPSC is considered to be proportional to N_{tot} the total number of synapses that receive stimulation that are release ready (R_{rel}), e.g. $N_{tot} \times R_{rel}$, multiplied by the release probability P_{rel}. Therefore, we can assume that the peak IPSC $\sim N_{tot} \times R_{rel} \times P_{rel}$,

where R_{rel} is the fraction of vesicles that are ready to release. Both R_{rel} and P_{rel} range between 0 and 1 and without loss of generality, we consider peak IPSC proportional to $P_{rel} \times R_{rel}$. The presynaptic calcium concentration Ca is assumed to follow first order decay kinetics relative to a base concentration, Ca_{base}. We assume that $Ca_{base} = 0$ because in the absence of action potential the concentration of calcium is fairly low. The evolution equation for Ca is

$$\tau_{ca} \frac{dCa}{dt} = -Ca, \tag{3.2.1}$$

where τ_{ca} is the calcium decay time constant in msec. Upon an action potential, calcium channels open and an influx of calcium ions into the synaptic terminal increase the concentration of calcium by an amount δ (measured in μm): $Ca \to Ca + \delta$ at the time of the pulse. The calcium concentration is scaled by the value of δ_c under control conditions, e.g, $C = \frac{Ca}{\delta_c}$. Following an action potential at a time $t = 0$, the calcium concentration C increases additively by an amount of $\Delta = \frac{\delta}{\delta_c}$, and therefore solution to the eqn. (3.2.1) is given by

$$C = C_0 e^{-t/\tau_{ca}} + \Delta. \tag{3.2.2}$$

Note that since $\Delta = 1$ in the control condition, eqn. (3.2.2) becomes

$$C = C_0 e^{-t/\tau_{ca}} + 1.$$

Following the work of Lee and colleagues [41], it is assumed that P_{rel} increases monotonically as function of calcium concentration in a sigmoidal fashion toward an asymptote value of P_{max}. The mechanism of vesicle binding and release depends upon a major calcium receptor (synaptotagmin-1) in the presynaptic terminal that binds the incoming calcium. This follows

a Hill equation with coefficient 4, and hence P_{rel} does so according to:

$$P_{rel} = P_{max}\frac{C^4}{C^4 + K^4},$$

where K is the half-height calcium concentration. Both K and P_{max} are parameters determined from the experimental data. It should be mentioned that P_{max} is calculated to be 0.87 in control condition and 0.27 in the muscarine condition through the mean-variance analysis [9, 40].

It is shown in [19, 65, 72] that the rate of recovery K_{recov} from refractory state depends on the calcium concentration in the presynaptic terminal and it follows a Hill equation with coefficient 1, starting at some k_{min}, increasing to k_{max} asymptotically as the calcium concentration increases, with a half height of K_r. Mathematically, the recovering rate is

$$k_{recov} = k_{min} + \Delta k\frac{C}{C + K_r},$$

where $\Delta k = k_{max} - k_{min}$.
The fraction of release-ready vesicles R_{rel} obeys the ordinary differential equation

$$\frac{dR_{rel}}{dt} = k_{recov}\left(1 - R_{rel}\right). \tag{3.2.3}$$

The differential equation (3.2.3) can be solved for R_{rel} and the solution is given by

$$R_{rel} = 1 - (1 - R_0)\left(\frac{C_0 e^{-t/\tau_{ca}} + K_r}{C_0 + K_r}\right)^{\Delta k} e^{-k_{min}t}. \tag{3.2.4}$$

Recall $IPSC \sim P_{rel} \times R_{rel}$, and upon vesicle release, R_{rel} is reduced by the fraction of synapses that fired results in $R_{rel} \to R_{rel} - P_{rel}R_{rel}$. This value, R_{rel}, can be assumed to be the initial condition for solution to the ODE eqn. (3.2.4). Given an interspike interval T, a

two dimensional map in C and R_{rel} that captures the peak value of the IPSC is given by

$$\begin{aligned} C_{n+1} &= C_n e^{-T/\tau_{ca}} + \Delta \\ Pr_{n+1} &= P_{max} \frac{C_{n+1}^4}{C_{n+1}^4 + K^4} \\ R_{n+1} &= 1 - (1 - (1 - P_n)\, R_n) \left(\frac{C_n e^{-T/\tau_{ca}} + K_r}{C_n + K_r} \right)^{\Delta k} e^{-k_{min} T}. \end{aligned} \tag{3.2.5}$$

Note that for the purpose of simplicity we set $Pr = P_{rel}$, $R = R_{rel}$, and normalized IPSC as P_r.

3.2.1 Parameter estimation

The parameter values are estimated by numerically minimizing the objective function (sum of the squared differences between predicted and observed values function) using the functions LSQNONLIN and Monte Carlo Markov Chain (MCMC) [27, 26]. These functions are available in the Matlab. The parameter description and fitted values are shown in the Tables 3.2.1 and 3.2.2.

Table 3.2.1: Parameters in the map given by eqn. (3.2.5)

Parameter	Description
Δ	Increase in the amount of calcium relative to that seen under control conditions
P_{max}	Maximum probability of release
K	Half calcium concentration value for probability of release function
k_{min}	Minimum rate of recovery of synapses
k_{max}	Maximum rate of recovery of synapses
K_r	Half calcium concentration value for rate of recovery function
τ_{ca}	Decay constant for calcium

It should be noted that $\Delta = 1$ under control condition, and $\Delta = 0.17$ in the muscarine condition.

Table 3.2.2: Fitted parameter values

Parameter	Fitted value
K	0.2
k_{min}	0.0017 1/msec
k_{max}	0.0517 1/msec
K_r	0.1
τ_{ca}	1.5 msec

3.2.2 Discussion of the model

A synapse can be classified to be depressing, facilitating or a mixture of the two, depending on its response to a train of action potentials. A number of models have been developed based on the condition that a synapse is either facilitating, depressing, or mixed, individually. This model, however, is built so that it combines both facilitating and depressing mechanisms. In other words, depending on the parameter values, facilitation, depression, and mixture of the two can be represented by this model.

3.3 Numerical Study of the Response (P_r) Distributions

Recall that postsynaptic responses are generated from the given map in (3.2.5). Assume that the interspike intervals (ISI's), identified by T in the map are generated from an exponential distribution with firing rates $\lambda > 0$. We are interested in exploring the effect of different firing rates on the postsynaptic response in muscarine and control conditions. Before presenting more details, it is important to understand the distribution of the data using a graphical method. In fact, John Tukey [69] recommended the practice of exploratory data analysis (EDA) as an essential part of the scientific process. We begin our analysis using histograms with equal bin width using the *Freedman and Diaconis Rule* introduced in section 2.3.2 of chapter 2. We generate 10^5 samples from P_r according to the map and discard initial transient. The relative frequency histograms of P_r for the firing rate in theta, gamma, and

higher (non-physiological, for comparison) under control conditions are shown in Figure 3.1. There is a clear difference in relative frequency distributions under stimulation at 0.5, 3, 8, 10, 20, and 100 Hz. At the low firing rates between 0.1 and 1 Hz, the relative frequency distributions are skewed to the left. Notice that for this range of firing rates the distribution is peaked near P_{max}. This is expected since the exponentially distributed ISI's with low firing rates between $0.1 - 1$ Hz contribute to refilling unavailable docking sites and hence the size of the readily releasable pool. This size is likely to play an important role in the probability of release at a synapse and so on increase in its P_r values. For very high firing rates (non-physical, 200 Hz and larger), the distributions are skewed to the right and peaked near very small values of P_r, almost 0, reflecting the fact that due to very small the ISIs, synapses do not have enough time to recover and return to a release-ready state. Therefore, variation in ISIs in the stimulus, as well as by the transmitter release contribute to the P_r distributions. As the firing rate increases from 0.5 to 10 Hz, the distribution of P_r becomes less severely skewed and more spread out over the entire interval. This is due to the range of firing rates presented in the exponential distribution of ISIs that results in more variation in responses. The coefficient of variation values (CV) for P_r's in Table 5.6.2 support these conclusions. In addition, the left-skewed P_r distribution obtained under very low firing rates shifts toward being flat or uniform at around 1.8 Hz before achieving the maximum variation at 3 Hz. Following this event, the peak of the distribution starts shifting to the left and variation slightly decreases. Note that 8 and 10 Hz cover the range of rates considered to be the theta rate. Here the synapse is the most sensitive, allowing for widely varying responses.

For firing rates between 20 and 100 Hz (the gamma range), the variation of response values is lower compared to the ones in the theta range and the peak sharpens on the left as the rate in increased. However this transition is slow and the peak near 0.1 persists for rates below 200 Hz and after which it is subsumed into the peak near 0. Therefore, it appears that the responses have distinct frequency "tunings" at physiological rates. From Table 5.6.2 and

Figure 3.1, it is easy to see that the histograms differ depending on the firing rate and that the mean and variance are reduced at higher firing rates.

Recall that when fitting the model to the muscarine data, a smaller value of Δ was needed ($\Delta = 0.17$). This situation is consistent with the assumption that muscarine shuts down the influx of calcium ions to the presynaptic terminal and hence reduces the size of response. This mechanism, however, reduces the relative amount of depression at firing rates around gamma which can have important implications for the effect of neuromodulation at these firing rates. It is important to investigate how this mechanism is revealed through the distribution of P_r. Figure 3.2 shows the relative frequency distribution of P_r under muscarine condition for varying firing rates. Similar pattens can be observed in the P_r distribution histograms under muscarine condition as were seen in control condition. However, as can be expected, since $P_{max} = 0.27$ in muscarine condition, P_r can take a range of values from 0 to 0.27 compared to control condition with $P_{max} = 0.87$. From Figure 3.2 and Table 5.6.3, it can be seen that for low firing rates between 0.1 and 0.5 Hz, the distribution is left skewed and peaked near P_{max}. At 3 Hz, the distribution is roughly symmetric and centered around 0.19. Although the peak of distribution moves gradually towards the left as firing rates increases, the symmetric shape of the distribution is preserved. Also, similar to control condition, as firing rate increases, the mean and variance of P_r decreases. Note that the mean and variance is smaller in muscarine conditions compared to that of the control conditions throughout the firing rates. Therefore, we conclude that muscarinic synapse focuses the response in a narrow interval centered around a small P_r values. It should be noted that in both conditions the dynamic of the map creates a low pass filter, which is an indication of a depressing synapse.

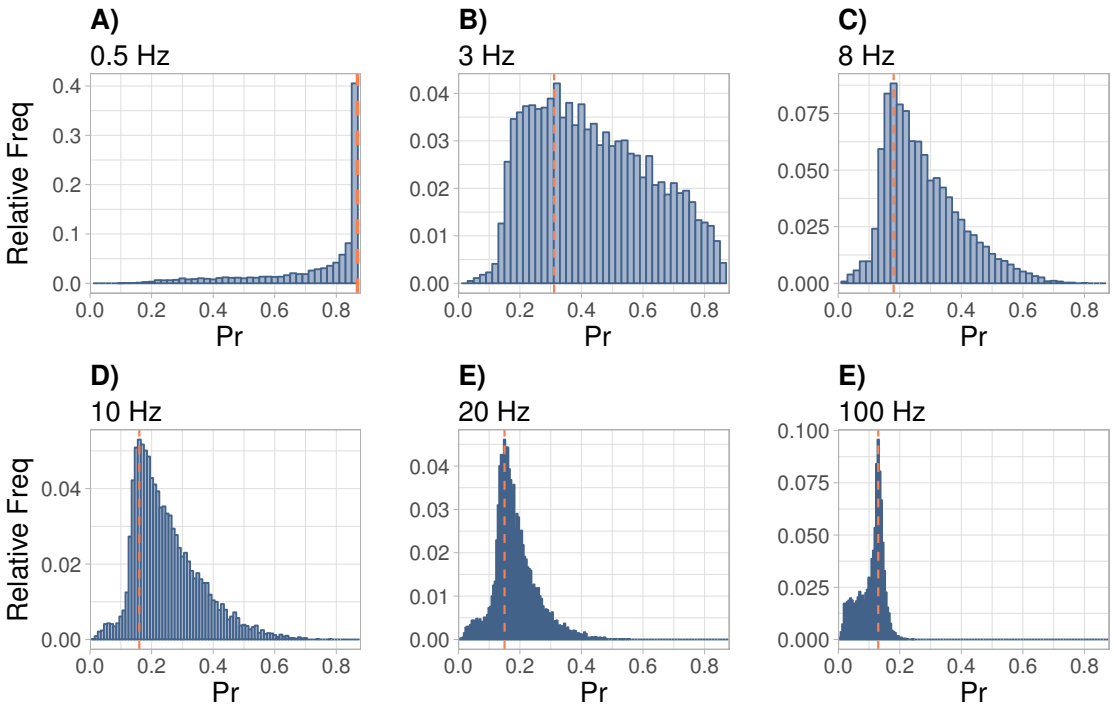


Figure 3.1: Estimated normalized response (P_r) distributions with the control parameter set under stimulation at firing rates A) 0.5, B) 3, C) 8, D) 10, E) 20, and F) 100 Hz. Horizontal axis shows the P_r values and the vertical axis is the relative frequency.

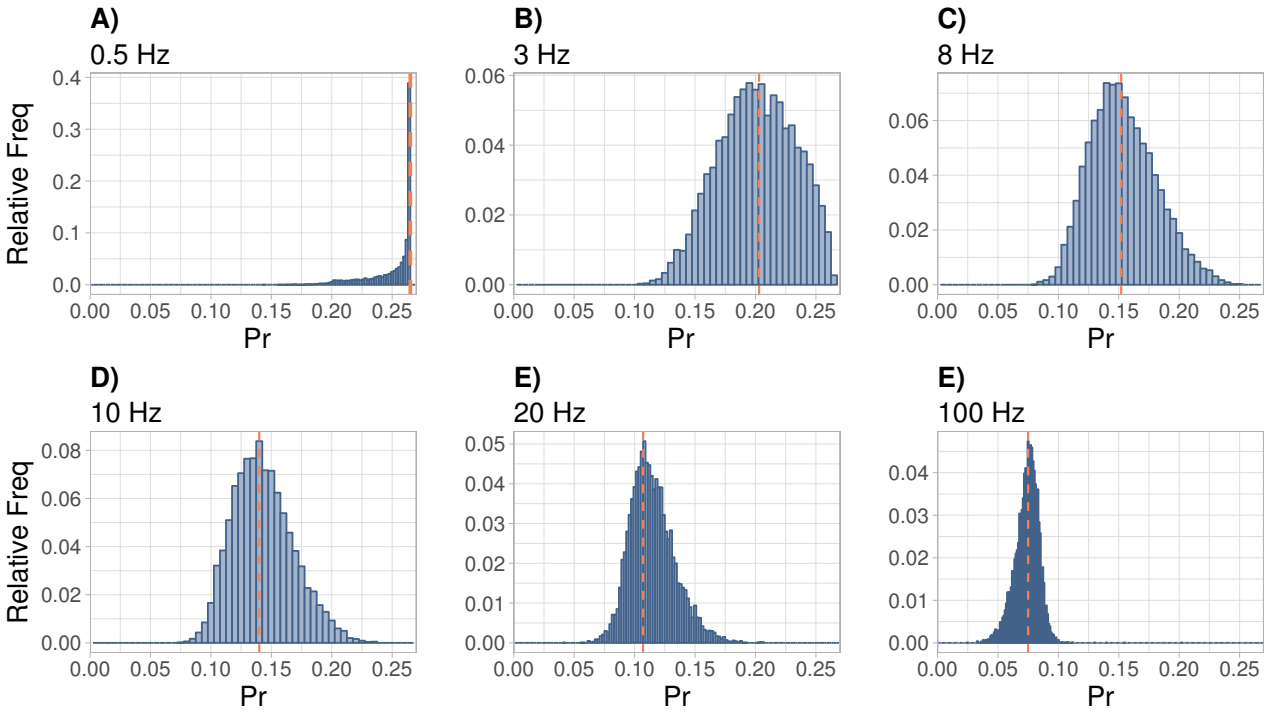


Figure 3.2: Normalized response (P_r) distribution with the muscarine parameter set under stimulation at firing rates A) 0.5, B) 3, C) 8, D) 10, E) 20, and F) 100 Hz. Horizontal axis shows the P_r values, vertical axis shows the relative frequency.

Table 3.3.1: Summary statistics of P_r's under control conditions.

	$\bar{x}$	s	CV	min	Q_1	M	Q_3	max
rate= 0.5 hz	0.736	0.178	0.242	0.108	0.664	0.826	0.866	0.868
rate= 3 hz	0.437	0.189	0.432	0.039	0.280	0.415	0.581	0.867
rate= 8 hz	0.275	0.122	0.445	0.008	0.181	0.249	0.346	0.803
rate= 10 hz	0.249	0.108	0.435	0.012	0.169	0.225	0.309	0.776
rate= 20 hz	0.184	0.069	0.377	0.006	0.142	0.172	0.217	0.556
rate= 100 hz	0.107	0.040	0.374	0.003	0.080	0.121	0.136	0.243

Table 3.3.2: Summary statistics of P_rs under muscarine conditions.

	$\bar{x}$	s	CV	min	Q_1	M	Q_3	max
rate= 0.5 hz	0.250	0.019	0.078	0.145	0.243	0.260	0.264	0.264
rate= 3 hz	0.200	0.031	0.156	0.103	0.178	0.201	0.225	0.264
rate= 8 hz	0.153	0.027	0.177	0.078	0.133	0.151	0.172	0.252
rate= 10 hz	0.143	0.025	0.175	0.074	0.125	0.141	0.159	0.236
rate= 20 hz	0.114	0.018	0.162	0.041	0.102	0.113	0.125	0.205
rate= 100 hz	0.074	0.010	0.135	0.024	0.068	0.075	0.080	0.202

We can compare the mean and peak of the P_r distribution for different firing rates under both muscarine and control conditions. It can be seen in Figures 3.3 and 3.4 that the mean and peak of the response distribution under control conditions are larger than those under muscarine conditions. However, within the theta range, the peak of the distributions under both control and muscarine conditions are very close, which can be interpreted that under depression caused by these pharmaceutical applications the synapse response remains stable at theta frequencies. Also, the amount of depression relative to the initial response of the postsynaptic neuron is larger under control conditions compared to muscarine: the mean ranges from 0.8 to roughly 0.1 under control condition, a change of about 0.7 overall, while under muscarine conditions, it ranges from 0.3 to 0.5; a significantly smaller range of .25. This confirms the assertion discussed in [66].

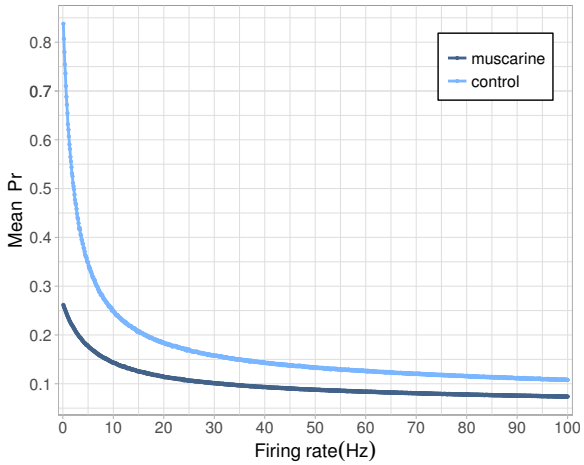


Figure 3.3: Mean of the normalized response P_r distribution plotted against firing rate.

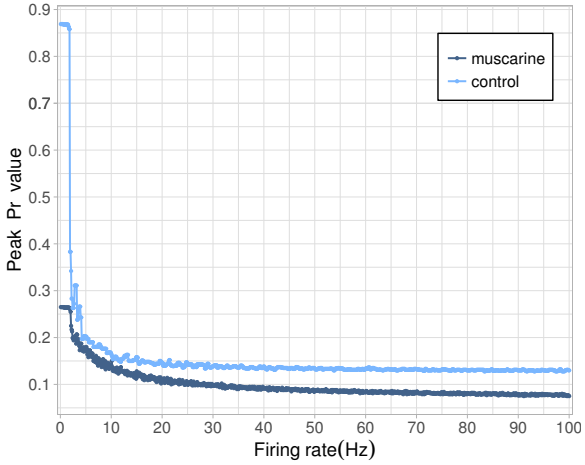


Figure 3.4: Peak of the normalized response P_r distribution plotted against firing rate.

3.4 Entropy of P_r vs. Mean Firing Rate

Recall from section 2.2, we introduced the notion of mutual information. In this section, we only specify some details of the information computation performed for deterministic models of depressing synapse in muscarine and control conditions. We estimate the information about the preceding interspike intervals (ISIs) contained in the postsynaptic response (P_r). The synapse is stimulated by the spike train S with particular mean firing rate, and we denote the associated series of interspike intervals by $I_S = \{t_1 - T_1, t_2 - t_1, t_3 - t_2, \cdots, t_N - t_{N-1}\}$, where T_1 is considered to be the beginning of the recorded trial.

We assume a popular model for spike trains, the homogeneous Poisson process, which is commonly used by the neuroscience community, because of its success in describing spike data recorded in vivo. Here we mention two properties which account for its adequacy.

- If we write $I_S = \{ISI_1, ISI_2, \cdots, ISI_{N-1}\}$, the we can consider the ISI's to be independent and identically distributed random variables.
- $f_{X=ISI}(x;\lambda) = \lambda e^{-\lambda x}$ has exponential distribution, where λ is referred to as the mean

firing rate for spike train, $\frac{1}{\lambda}$ is the mean ISI of the process.

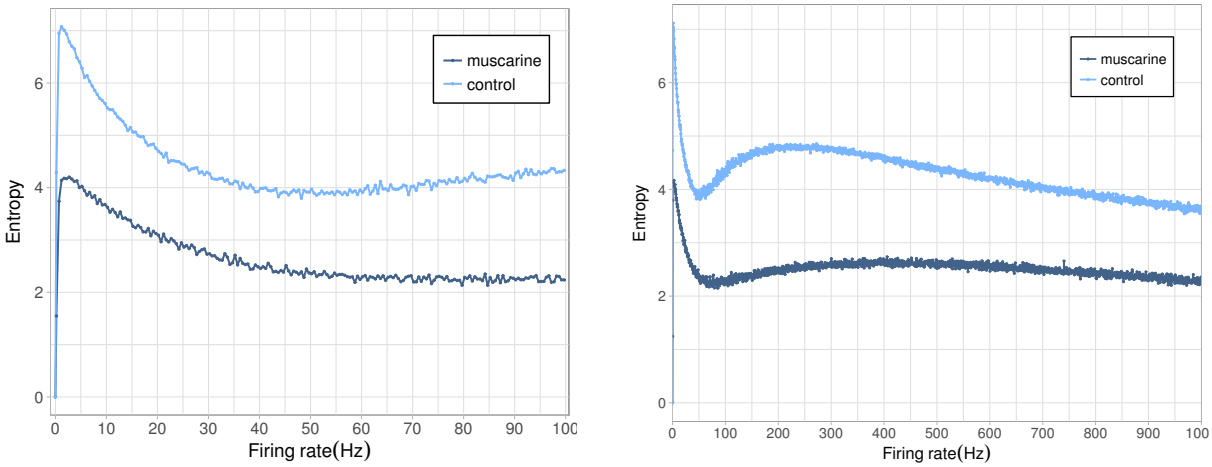
The estimated mutual information between P_r and the $ISIs$ is obtained from $\hat{I}(P_r; ISIs) = \hat{H}(P_r) - \hat{H}(P_r|ISIs)$. However, because the synapse is deterministically modeled, the postsynaptic response is uniquely determined by the sequence of preceding interspike intervals. Recall from subsection 2.2.5, if P_r is completely defined by $ISIs$, then $I(P_r; ISIs) = H(P_r)$. Thus, the information contained in P_r about the preceding $ISIs$ is solely given by the unconditional entropy of P_r distribution. The P_r variables are discretized using bin sizes obtained by applying the *FD Rule*, which was presented in detail in chapter 2 subsection 2.3.1.
Figure 3.5 illustrates the entropy of the postsynaptic response as a function of the mean firing rate obtained from exponentially distributed $ISIs$ for both control and muscarine parameter sets. Figure 3.5 (b) aims to show the behavior of information entropy in the limit for large mean firing rates ranging from near 0 to 1000 Hz, while Figure 3.5 (a) zooms in on the physiological range of mean firing rates between 0 and 100 Hz.

Within range of 0.1 to 100 Hz mean firing rates, there is a local maximum in entropy for both control and muscarine conditions. This maximum entropy for both conditions occurs between 1 and 4 Hz. In this range $(0.1 - 100)$, the P_r distribution is spread out between a peaked P_r at high values to the one at lower values, with large variability in the size of response. Under control conditions, for larger firing rates (non- physiological) the entropy increases again to a local maximum near 200 Hz, followed by decay as the P_r distribution declines to zero. Under muscarine conditions, however, the second local maximum occurs around 400 Hz.

Under both conditions, the mean and variance of the P_r distributions are reduced at higher firing rates, resulting in reduction in the entropy of postsynaptic response P_r distribution. This is expected, since variance and information are interchangeable in the context of Shannon information theory and thus entropy increases when variation increases. [46, 7]. Therefore, it

is likely that the deterministic model of synapse behavior reflects less information at higher firing rates.

We should mention that the size of postsynaptic response is a measure of strength in the synaptic connection and so having a skewed or narrowed distribution peaked at higher values of P_r leads to a stable synaptic connection, even when presented with a stochastic signal of the Poisson type. Alternatively, distributional transitions from higher peaked values to lower ones as the mean firing rate increases within the theta range, the entropy is maximized and hence greater range of coupling strengths is created. The value of this strength depends on the presynaptic activity of neuron.



(a) (b)

Figure 3.5: Entropy of the normalized response P_r when stimulated with Poisson distributed spike trains of varying mean firing rate for the control and muscarine parameter sets.

3.5 The Stochastic Model of the Postsynaptic Response

In this section, we expand the model in section 3.2 introduced by Stone and her colleagues [66] to a stochastic version in which individual vesicles available in release sites are released probabilistically. We then measure the correlation between postsynaptic response PSR and preceding $ISIs$ and estimate the amount of information being transferred using information theoretic functionals following the approach of Fuhrmann [23].

3.5.1 Model description

Noise is a fundamental constraint to information transmission. In 1920s, Adrian who was one of the first to record from neurons noticed that neuronal responses were highly variable across identical stimulation trials and only the average of responses could be related to the give stimulus. [2, 3, 49]. Biologists refer to it as "Variability", while engineers commonly call it "Noise". Therefore, the brain processes information in the presence of variability and so it is natural to consider a model that takes into account the neuronal trial-to-trial variability. Researchers established that one of the sources of this variability is the stochastic release of neurotransmitter into synaptic cleft. Here, we follow the work of Katz and his colleagues to describe the physiological mechanism of neurotransmitter release [64].

Upon the arrival of an action potential at a presynaptic terminal, the influx of calcium through calcium channels leads to the fusion of some vesicles with the axon terminal membrane at special release sites. This leads to the release of neurotransmitters into the synaptic cleft. Each release site can release either one or zero vesicles. We denote the number of release sites by N_{tot}. Each event, i.e., a release of neurotransmitter, occurs independently of all others. This mechanism can be likened to flipping a coin at each release site and if the

coin lands head side up, the vesicle releases, otherwise no release occurs. Therefore, we can assume that release of K vesicles from N_{tot} release sites follows a Binomial distribution with two parameters (N_{tot}, P_r), and write $K \sim B(N_{tot}, P_r)$. The PDF is

$$P\left(k; N_{tot}, P_r\right) = P\left(K = k\right) = \binom{N_{tot}}{k} P_r^k (1 - P_r)^{N_{tot}-k},$$

for $k = 0, 1, \cdots, N_{tot}$, where $\binom{N_{tot}}{k} = \frac{N_{tot}!}{k!(N_{tot}-k)!}$ is the binomial coefficient and P_r is the release probability for each site following an action potential. Note that the variability is not only due to the probabilistic nature of the number of vesicles being released, but also in the postsynaptic response to a single vesicle. This is due to many factors such as variation in the number neurotransmitters contained in synaptic vesicles or variation in receptor binding [24]. Thus, we assume that the size of the postsynaptic response (Q_{rep}) at the time of the spike is not a constant, but it follows Normal distribution with parameters mean μ and standard deviation σ, with a two-sided truncation, and write $Q_{resp} \sim N(\mu, \sigma)$. Note that failure of a release results in a zero amplitude response from the postsynaptic neuron, thus it cannot be informative.

Therefore, the amplitude of postsynaptic response following each action potential is obtained by combining the Binomial model of vesicle release with Normal model of a single response. Hence the summation of responses evoked by each vesicle release is given by

$$Q_{resp} = \begin{cases} 0 & \text{if } K = 0, \\ \sum_{i=1}^{k} Q_{resp_i} & \text{if } K > 0, \end{cases}$$

where K is the number vesicle that are released from the total number of release sites. Note that for $K > 0$, $Q_{resp} \sim N(K\mu, \sqrt{K}\sigma)$. Therefore the probability density function of the

postsynaptic response to release of single vesicle is given

$$f(Q_{resp} = q|\mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}\mathcal{N}_{cts}} e^{-\frac{(q-\mu)^2}{2\sigma^2}} & \text{if } 0 < q < 2\mu, \\ 0 & \text{if O.W.} \end{cases} \tag{3.5.1}$$

Where

$$\mathcal{N}_{cts} = \int_0^{2\mu} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(q-\mu)^2}{2\sigma^2}} dq.$$

is a normalizing factor. We will use this formulation in what follows.

3.5.2 Mutual information calculations

We use Shannon information theory to estimate the mutual information content in the postsynaptic response PSR, denoted by Q_{resp} in subsection 3.5.1, about the preceding interspike intervals $ISIs$ in both muscarine and control conditions in the stochastic model of a synapse. We stimulated the synapse with Poisson spike trains with varying mean firing rates. Therefore, the estimated mutual information is given by

$$\hat{I}(PSR; I_S) = H(PSR) + H(I_S) - H(PSR, I_S) \tag{3.5.2}$$

For more details on calculation of mutual information, see chapter 2 subsection 2.3.1.

Figure 3.6 shows the results for the control and muscarine parameter sets. Under both conditions, we observe a rapid decline in information transfer at stimulation firing rates above a few Hertz, and maximal transfer which occurs between 0.1 to 2 Hz. This demonstrates the frequency dependence of temporal information encoding. Larger interspike intervals (lower mean firing rate) allow sufficient time for the release sites to recover, leading to a narrow and skewed distribution of postsynaptic responses peaked around P_{max}, and hence the communication of information is very limited. In contrast, there is not enough time in very short interspike

intervals (high firing rate) for the synapse to recover and all the responses are depressed and accumulate near zero, leading to low information transmission. Between these two extremes, interspike intervals have a significant effect on the amplitude of PSR. The major difference between muscarine and control conditions is in the absolute value of the mutual information, which is significantly lower under muscarine conditions. This is expected because the values that PSR can take under muscarine conditions are smaller than those under control conditions. This agrees with the results in [23], where mutual information in depressing synapses is maximized at low firing rates.

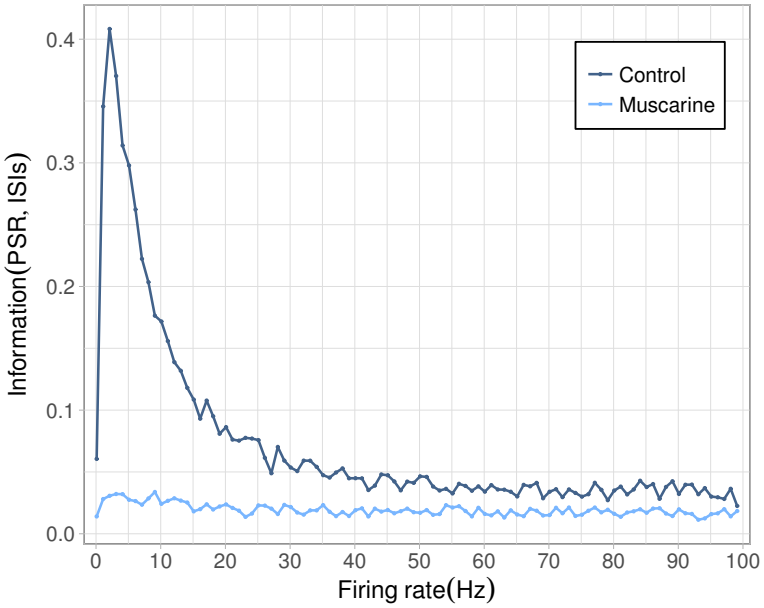


Figure 3.6: Estimated mutual information transmission between normalized postsynaptic response PSR and exponentially distributed preceding $ISIs$ with mean firing rates ranges from 0.1 to 200 Hz.

3.6 Information "Stored" in the Postsynaptic Response

In previous section we showed that the dynamics of a synapse stores information originating from the temporal structure of the preceding presynaptic activity such as spike timing or

interspike intervals. This information is transmitted to the postsynaptic neuron through the amplitude of the postsynaptic response. Therefore, we can assume that a synapse serves as a "transit memory buffer" that stores and transmits information. It is clear that a synapse can carry information about a finite number of preceding spike times. Our interest is in estimating this quantity. Markram and Mass in 2002 [43] drew an interesting analogy by comparing the information storage in a synapse about the preceding spike train to hashing in computer science, where items from a large universe are assigned to addresses in s memory structures with smaller number of slots than the number of items in universe V [14]. In our case, the universe is all possible preceding spike trains and the slots in the memory structure correspond to the dynamic states of the synapse.

Considering this, we estimate the amount of information contained in normalized postsynaptic response (PSR) about the sequential number of the preceding presynaptic spikes. Here we present two approaches. In the first approach, we compute the mutual information between postsynaptic response and the summation of k preceding interspike intervals, for $k = 1, \cdots, m$. This approach, i.e. using single summary measurement to reflect the temporal activity of presynaptic neuron, however, might be biased because a presynaptic neuron with the same measured value may have a distinct combination of $ISIs$ that effects the PSR differently and thus this single measurement ignores the subtleties of the exact sequence. For instance, a relatively long ISI followed by a short ISI results in a PSR that is smaller than the reverse, although the sum of two is the same. The second approach is to compute mutual information between PSR and an $m - tuple$ of preceding interspike intervals where m is an integer.

Information between postsynaptic response PSR and sum of preceding $ISIs$

We compute the mutual information between postsynaptic response and a sum of the preceding $ISIs$. A larger number of terms in the sum indicates that the spikes that occurred further back in time. Let the beginning of the recorded trial be the time when the first spike occurs ($t_1 = 0$). Let $\mathbb{T}_1 = ISI_{(1)}, \mathbb{T}_2 = ISI_{(2)} + ISI_{(1)}, \mathbb{T}_3 = ISI_{(3)} + ISI_{(2)} + ISI_{(1)}, \cdots, \mathbb{T}_k =$

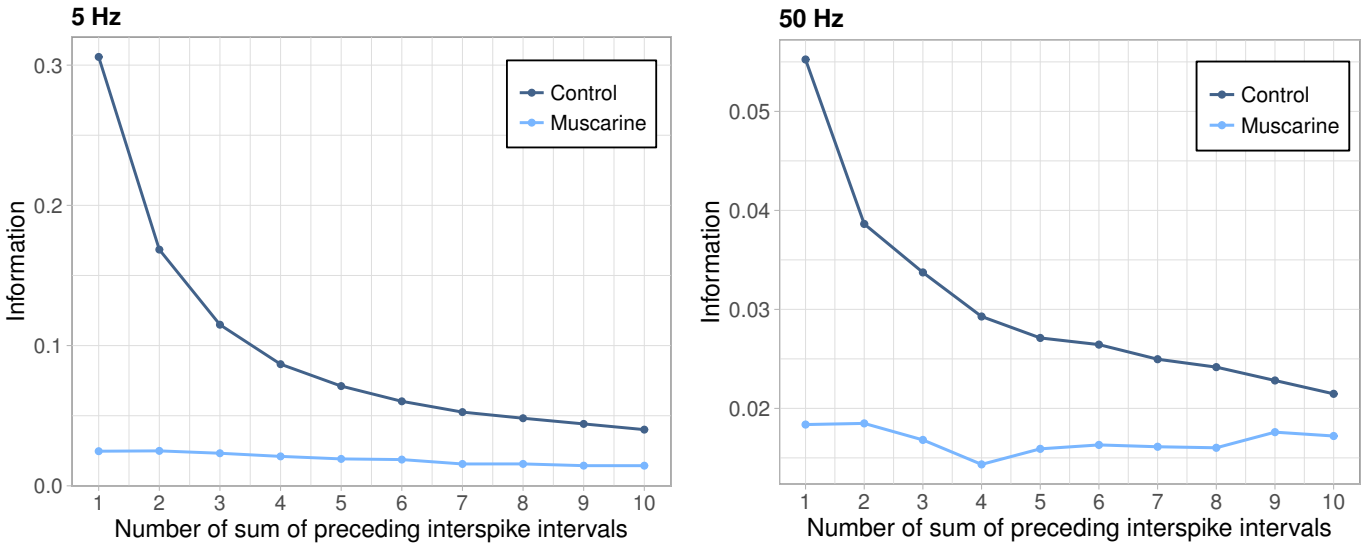
$ISI_{(k)} \cdots + ISI_{(2)} + ISI_{(1)}$ be a vector of the sum of the m preceding $ISIs$, in order. It can be simplified as

$$\mathbb{T}_k = \sum_{i=1}^{k} ISI_{(i)}, \qquad\qquad k = 1, \cdots, m,$$

where $m > 0$ is a natural number. Thus, the estimated mutual information between postsynaptic response and the m preceding interspike intervals is given by

$$\hat{I}\left(PSR; \mathbb{T}_k\right) = \hat{H}\left(PSR\right) - \hat{H}\left(PSR|\mathbb{T}_k\right), \qquad \text{for } k = 1, \cdots, m. \tag{3.6.1}$$

In Figure 3.7, the information content in PSR is plotted against the sequential number of preceding presynaptic interspike intervals for muscarine and control conditions, at 5 and 50 Hz. It comes as no surprise, given the result from preceding sections, that the information content is significantly lower in muscarine compared to control condition. For both firing rates, 5 and 50 Hz, in the control condition the estimated mutual information decreases as more ISI terms are included in the sum. Hence the further back in time the sum goes, the less the past terms ISI are directly involved in determining PSR. It can be seen that in control conditions for 5 and 50 Hz, the PSR carries information about a sum of almost 4 preceding $ISIs$. This shows that depressing synapse can encode information about the sum of 4 preceding interspike intervals. Also, mutual information in muscatine condition is less dependent on the cumulative history of spikes as their consecutive terms proportions, $\frac{ISI_{k-1}}{ISI_k}$, is almost constant and indicates hardly any changes as m increases. The reason for this is simple. In the muscarine case the range of postsynaptic response range is narrow, and thus cannot carry as much information from preceding $ISIs$.



(a) (b)

Figure 3.7: Information between the postsynaptic response and the preceding sums of interspike intervals in control and muscarine cases.

3.6.1 Mutual information between normalized postsynaptic response PSR and m-tuple inter-spike intervals $ISI_1, ISI_2, \cdots, ISI_m$

The second approach preserves the exact structure of the sequence of the preceding $ISIs$ but is much more computationally intensive.

Theorem. *: [Chain rule for entropy][15]*

Let $X_1, X_2, \cdots, X_m$ be drawn according to $P(x_1, x_2, \cdots, x_m)$. Then,

$$H(X_1, X_2, \cdots, X_m) = \sum_{i=1}^{m} H(X_i | X_{i-1}, \cdots, X_1). \tag{3.6.2}$$

We can rewrite eqn. (3.6.2)

$$\begin{aligned} H\left(X_1, X_2, \cdots, X_m, Y\right) &= \sum_{i=1}^{m+1} H\left(X_i | X_{i-1}, \cdots, X_1\right) \\ &= H\left(Y | X_m, \cdots, X_1\right) + \sum_{i=1}^{m} H\left(X_i | X_{i-1}, \cdots, X_1\right) \\ H\left(Y | X_1, \cdots, X_m\right) &= H\left(X_1, \cdots, X_m, Y\right) - H\left(X_1, \cdots, X_m\right), \end{aligned} \tag{3.6.3}$$

where denote $X_{m+1} = Y$.

As well as for entropy, a chain rule for information can be formulated.

Theorem. *: [Chain rule for information][15]*

Let $X_1, X_2, \cdots, X_m$ be drawn according to $P(x_1, x_2, \cdots, x_m)$. Then,

$$\begin{aligned} I\left(X_1, X_2, \cdots, X_m; Y\right) &= \sum_{i=1}^{m} I\left(X_i; Y | X_{i-1}, \cdots, X_1\right) \\ &= H\left(Y\right) - H\left(Y | X_1, X_2, \cdots, X_m\right). \end{aligned} \tag{3.6.4}$$

For notational simplicity we will denote ISI with X and PSR with Y. Consider a single-input (SI) and single-output (SO) channel with input interspike interval X and output postsynaptic response Y. The amount of transmitted information between a postsynaptic response and preceding interspike interval can be simply calculated from eqn. (3.6.4) and is given by

$$\begin{aligned} I\left(Y; X_1\right) &= H(Y) - H(Y | X_1) \\ &= H(Y) + H(X_1) - H(X_1, Y). \end{aligned} \tag{3.6.5}$$

Now consider a channel with two inputs preceding interspike intervals X_1 and X_2 and a single output response Y. We can define the transmitted information $I(< X_1, X_2 >; Y)$ between the

output and two inputs variables X_1 and X_2 jointly as

$$I\left(Y;< X_1, X_2 >\right) = H(Y) - H(Y|X_1, X_2). \tag{3.6.6}$$

The information can be rewritten by applying eqns. (3.6.2) and (3.6.4) to eqn. (3.6.6) to obtain

$$\begin{aligned} I\left(Y;< X_1, X_2 >\right) &= H(Y) - H(X_1, X_2, Y) + H(X_1) + H(X_2|X_1) \\ &= H(Y) - H(X_1, X_2, Y) + H(X_1) + H(X_1, X_2) - H(X_1) \\ &= H(Y) + H(X_1, X_2) - H(X_1, X_2, Y). \end{aligned} \tag{3.6.7}$$

We refer to this measure as the mutual information between response Y and the tuple $< X_1, X_2 >$, and it can be interpreted as the reduction of uncertainty in the response Y due to knowledge of $< X_1, X_2 >$.

Following McGill [45] we can extend the definition for mutual information to include more than two preceding interspike intervals $X_1, X_2, \cdots, X_m$ that transmit to postsynaptic response Y as follows

$$\begin{aligned} I(< X_1, X_2, \cdots, X_m >; Y) &= H(Y) - H(Y|X_1, X_2, \cdots, X_m) \\ &= H(Y) + H(X_1, X_2, \cdots, X_m) - H(X_1, X_2, \cdots, X_m, Y). \end{aligned}$$

Note that it is difficult to reliably estimate joint probability distributions of high dimensionality. The estimate of $f(x_1)$ is more robust than the estimate of $f(x_1, x_2, \cdots, x_m)$. As we mentioned in the previous chapters, the most common and well-known method of calculating mutual information from empirical data is to use binning to create an approximate probability density distribution. However, this method is prone to bias in higher dimensions. Therefore, we use the KSG algorithm instead to estimate mutual information between the postsynaptic response Y (PSR) as single-output and m preceding interspike intervals $ISIs$ $(X_1, \cdots, X_m)$

as a multivariate inputs.

In order to do this, we need to introduce yet another notion which is "Total Correlation". In 1960 Watanable [73] was one of the first to discuss total correlation in detail, although the same concept had been described previously by McGill [45]. Palus [50] and Weinholt and Sendhoff [75] refer to it as "Redundancy", while Tononi [68] called this measure "Neural Complexity (CN)". Total correlation is sometimes referred to as "multi-information" or "multivariate mutual information" [6].

Definition 3.6.1. [45] Multi-information or total correlation or redundancy among a $m+1$ set of random variables, $X_1, X_2, \cdots, X_m, Y$ is defined as:

$$R(X_1; X_2; \cdots; X_m; Y) = \sum_{i=1}^{m} H(X_i) + H(Y) - H(X_1, X_2, \cdots, X_m, Y).$$

This measure is symmetric, non-negative and non-decreasing with the number of variables. For $m = 1$ this measure boils down to the mutual information,

$$R(X_1; Y) = I(X_1; Y).$$

For $m = 2$,

$$R(X_1; X_2; Y) = H(X_1) + H(X_2) + H(Y) - H(X_1, X_2, Y).$$

If we add and subtract $H(X_1, X_2)$, we obtain

$$\begin{aligned} R(X_1; X_2; Y) &= H(X_1) + H(X_2) + H(Y) - H(X_1, X_2, Y) \\ &= H(X_1) + H(X_2) - H(X_1, X_2) + H(X_1, X_2) + H(Y) - H(X_1, X_2, Y) \\ &= I(X_1; X_2) + I(< X_1, X_2 >; Y). \end{aligned}$$

Similarly when $m = 3$, by adding and subtracting $H(X_1, X_2, X_3)$ we have

$$\begin{aligned} R(X_1; X_2; X_3; Y) &= H(X_1) + H(X_2) + H(X_3) + H(Y) - H(X_1, X_2, X_3, Y) \\ &= H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3) + H(X_1, X_2, X_3) + H(Y) - \\ & H(X_1, X_2, X_3, Y) \\ &= R(X_1; X_2; X_3) + I(< X_1, X_2, X_3 >; Y). \end{aligned}$$

This can be extended to m+1-dimension as

$$R(X_1; X_2; \cdots; X_m; Y) = R(X_1; X_2; \cdots; X_m) + I(< X_1, X_2, \cdots, X_m >; Y). \tag{3.6.8}$$

Note that we can simply rewrite eqn. (3.6.8) as

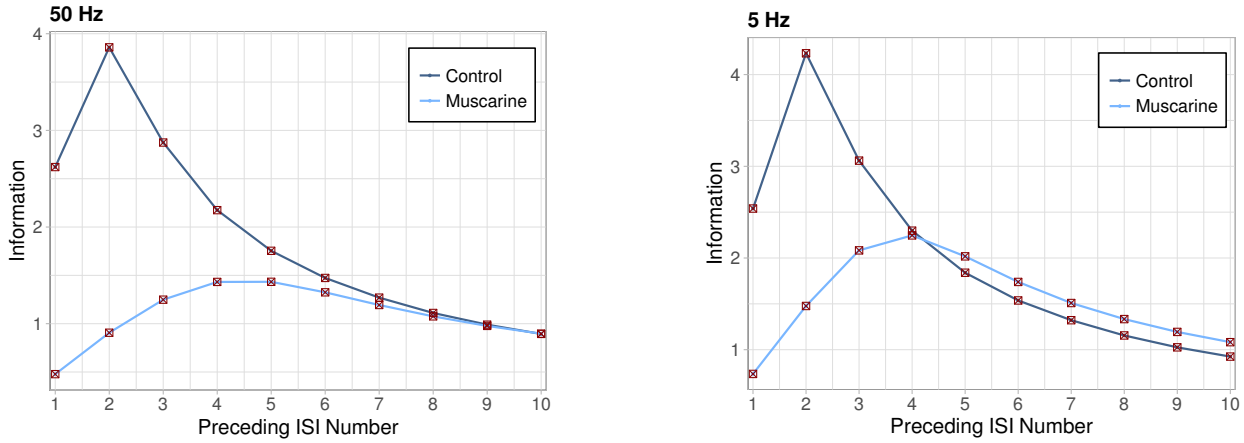
$$I(< X_1, X_2, \cdots, X_m >; Y) = R(X_1; X_2; \cdots; X_m; Y) - R(X_1; X_2; \cdots; X_m).$$

As we have seen in the preceding sections, Kraskov [39] gives formulas for generalized redundancies in higher dimensions. We use [39] to estimate the mutual information between $X_1, X_2, \cdots, X_m$ preceding interspike intervals and the postsynaptic response Y as follows

$$\hat{R}(X_1, X_2, \cdots, X_m, Y) = \psi(k) + m\psi(n) - < \psi(n_{x_1}) + \psi(n_{x_2}) + \cdots + \psi(n_{x_m}) + \psi(n_y) > .$$

Figure 3.8 shows the changes in the information (or reduction in uncertainty) between the postsynaptic response and m-tuple interspike intervals for increasing m. Mean firing rates

in the gamma and theta ranges of 50 and 5 Hz, respectively, are plotted for the control and muscarine conditions. It can be seen that at 50 Hz under both conditions the synapse is capable of memorizing about 5 preceding interspike intervals. At 50 Hz information in muscarine condition is always smaller than the control condition. Similarly, at 5 Hz this size is around 4 preceding interspike intervals for both control and muscarine conditions.



(a) (b)

Figure 3.8: N-tuple information between inter-spike intervals and postsynaptic response for control and muscarine cases for two frequencies: 5 and 50 Hz.

3.7 Discussion

When presented with Poisson spike train inputs. We observed that depressing synapse acts as a nonlinear filter for interspike intervals. For high and low mean firing rates, the output response distribution achieves a maximum at values near zero and P_{max}, respectively, aside a small degree of variation. However, for frequencies near theta, the response distribution is more spread across the entire interval between 0 and P_{max}. Over the gamma frequency range, the mode and the mean of the distribution remain almost constant, indicating a stable response size when presented with Poisson spike trains. This would create a stable connection and is the advantage of having a distribution with low entropy.

Given a deterministic map, we observed that the mutual information between postsynaptic response and the preceding interspike intervals is equivalent to the entropy of postsynaptic response as the response is uniquely determined by the preceding interspike intervals. In the stochastic model analysis, however, binomial variation was introduced due to release of neurotransmitter, and Gaussian variation due to fluctuation in response to a single vesicle release. We saw that the mutual information as a function of mean firing rate has a peak around 3 Hz for both the muscarine and control conditions, though the overall mutual information was much lower for muscarine condition. It was seen that the synapses with muscarine added as a neuromadulator are much less sensitive to changes in frequency overall.

We took our calculations a step further by investigating latency in synaptic response or the amount of temporal activity of presynaptic neuron synapses can store and transmit. To address this question, we calculated the mutual information between postsynaptic responses and the sum of previous interspike intervals, including successively more past times. This approach somehow ignores the structure of preceding temporal activity, measuring only if it was overall a long time or a short time, compared to other samples. We saw an overall decline in mutual information as more interspike intervals were added to the sum under control condition, while muscarine condition was somehow insensitive to this addition, being relatively flat.

In an attempt to maintain the structure in a preceding interspike intervals, we performed m-tuple information analysis using the KSG algorithm. This analysis measures the mutual information between the response and the sequence of interspike intervals as the number of interspike intervals in the sequence is increased. It was seen that if information increases with more interspike intervals in the sequence, then the reduction of uncertainty is greater as longer

histories are incorporated. It was seen that adding up to 5 interspike intervals to the sequence improves the prediction of P_r in control vs. 4 ISIs in muscarine. This result is more or less the same at gamma and theta frequencies. Note that using the sum of the ISIs in the mutual information calculation obsecures this dependence in the muscarine condition.

Through the analysis shown in this chapter, we have gained insight into the information processing characteristics of parvalbumin basket cells. It is clear from our quantitative description that constraining the firing of PV BCs at theta frequency optimizes the information content of PV BC to pyramidal cells. In the absence of muscarine, PV BCs are optimally tuned to transfer information at theta frequency. In the presence of cholinergic neuromodulation, when PV BCs may become depolarized and are more likely to fire at gamma frequency, the information processing capability is reduced.

Future studies include understanding the interaction of the synaptic dynamics with voltage oscillations in the hippocampus. In addition, analyzing the ability of synapses in information transmission in the presence of many presynaptic neurons influence on the postsynaptic cell in hippocampus is of interest. We will continue our quest to quantify the information processing properties of synapses by expanding beyond mutual information analysis.

Chapter 4

Parsimonious Approximate Descriptions of the Data

4.1 Motivation

We live in a complex world and its complexity makes it useful to consider that a sample obtained from measurements is coming from some probability distribution. The natural and more convenient approach to model this distribution with relatively small number of parameters is to use a parametric function. The parameters are then estimated using the sample. However, empirical data resulting from observed situations almost never exactly follow theoretical distributions. Nevertheless, empirical distributions are often similar to theoretical distributions. It is clear that using a specific parametric distribution can be useful as a model of the data. Therefore, the goal here is to obtain parsimonious approximate descriptions of the data. In other words, we would like to find the simplest viable distributional model for the data. In Chapter 3 it was shown that the map for calcium concentration C, the fraction of readily releasable vesicles R, and the release probability P driven by a Poisson spike train is

a random recurrence equation. In this chapter we show that while it is not possible to find a closed form for the exact map, an approximation can be created that reflect on the properties of the deterministic map.

4.2 Modeling Calcium Concentration Data with a Gamma Distribution

In this section, we show how the Gamma distribution can be applied to the calcium concentration data obtained from the two-dimensional map introduced in Chapter 3 section 3.2.

4.2.1 Characterization of the Calcium Distribution From a Random Difference Equation

A random difference equation (RDE) is an example of a random recurrence equation. The recursion is defined by a random affine linear function $\Psi(x) = Mx + Q$ for a pair (M, Q) of real-valued random variables. More precisely, let $(M_n, Q_n)_{n\geq 1}$ be a sequence of independent random variables with the same distribution as (M, Q), and define $(X_n)_{n\geq 0}$ recursively by

$$X_n = M_n X_{n-1} + Q_n, \qquad n \geq 1 \tag{4.2.1}$$

This is a general form of one-dimensional RDE and has been used in many applications to model a process that is subject to trend with respect to a variable n. An extensive discussion of the random difference equation is given in Vervaat [70] and Embrechts [21].

From eqn. (4.2.1) and by substitution, we obtain for $n = 1, 2, \cdots$

$$\begin{aligned} X_n &= M_n X_{n-1} + Q_n, \\ &= M_n M_{n-1} X_{n-2} + M_n Q_{n-1} + Q_n, \\ &= M_n M_{n-1} M_{n-2} X_{n-3} + M_n M_{n-1} Q_{n-2} + M_n Q_{n-1} + Q_n, \\ &\vdots \\ &= M_n M_{n-1} \cdots M_1 X_0 + \sum_{k=1}^{n} M_n \cdots M_{k+1} Q_k. \end{aligned}$$

Now use the independence assumptions and replace $(M_k, Q_k)_{1 \le k \le n}$ with the copy $(M_{n+1-k}, Q_{n+1-k})_{1 \le k \le n}$ to see that

$$X_n \stackrel{d}{=} M_1 M_2 \cdots M_n X_0 + \sum_{k=1}^{n} M_1 \cdots M_{k-1} Q_k, \tag{4.2.2}$$

for any $n \ge 1$, where $\stackrel{d}{=}$ denotes equality in distribution.

We are interested in the conditions that ensure the convergence in distribution of X_n. A fundamental theoretical result attributed to Kesten [34] states that if

$$\mathbb{E}\left(\ln |M|\right) < 0 \qquad \text{and} \qquad \mathbb{E}\left(\ln |Q|\right) < \infty \tag{4.2.3}$$

the series $\sum_{k=1}^{\infty} M_1 \cdots M_{k-1} Q_k$ will converge with probability 1. Then the sequence X_n converges in distribution to a random variable X, which necessarily satisfies the distributional identity

$$X \stackrel{d}{=} MX + Q.$$

See [70, 31] for more details on the convergence properties of X_n.

4.2.2 Random Difference Equation for Calcium Concentration

Suppose independent, identically distributed changes in the amount of calcium $\{\Delta_n\}_{n\geq 1}$ occur at times $\{t_n\}_{n\geq 1}$, and we are interested in the distribution of the amount of calcium accumulated C. In one dimension, the random recurrence equation for the calcium concentration is given by

$$C_n = A_n C_{n-1} + \Delta_n, \qquad n \geq 1, \tag{4.2.4}$$

where $A_n = e^{-T_n/\tau_{ca}}$, and the waiting times $T_1 = t_1$, $T_n = t_n - t_{n-1}$, $n \geq 2$, are i.i.d., making the $\{t_n\}$ a renewal process. Moreover, (A_n, Δ_n) are assumed to be i.i.d. vectors. C_0 is the base calcium concentration which is assumed to be zero.
Iterating (4.2.4) leads to

$$\begin{aligned} C_n &= A_n C_{n-1} + \Delta_n, \\ &= A_n A_{n-1} C_{n-2} + A_n \Delta_{n-1} + \Delta_n, \\ &= A_n A_{n-1} A_{n-2} C_{n-3} + A_n A_{n-1} \Delta_{n-2} + A_n \Delta_{n-1} + \Delta_n, \\ &\vdots \\ &= A_n A_{n-1} . \cdots . A_1 C_0 + \sum_{k=1}^{n} A_n . \cdots . A_{k+1} \Delta_k, \end{aligned}$$

for each $n \geq 1$. Using the independence assumptions and replacing $(A_k, \Delta_k)_{1\leq k\leq n}$ with the copy $(A_{n+1-k}, \Delta_{n+1-k})_{1\leq k\leq n}$ we observe that

$$C_n \stackrel{d}{=} A_1 A_2 . \cdots . A_n C_0 + \sum_{k=1}^{n} A_1 A_2 . \cdots . A_{k-1} \Delta_k.$$

Note that we assumed $C_0 = 0$, so

$$C_n \stackrel{d}{=} \sum_{k=1}^{n} A_1 A_2 . \cdots . A_{k-1} \Delta_k.$$

In previous section we indicated that based on the fundamental theoretical result [34] and given

$$\mathbb{E}\left(\ln|A|\right) < 0 \quad \text{and} \quad \mathbb{E}\left(\ln|\Delta|\right) < \infty$$

then the series

$$C = \sum_{k=1}^{\infty} A_1 A_2 . \cdots . A_{k-1} \Delta_k,$$

will converge with probability 1 and the distribution of C_n converges to that of C. Here we check these conditions to ensure the convergence in distribution of C_n. The first condition to check is that $\mathbb{E}\left(\ln|A|\right) < 0$. To verify, we calculate

$$\begin{aligned}
\mathbb{E}\left(\ln|A|\right) &= \mathbb{E}\left(\ln|e^{-T/\tau_{ca}}|\right), \\
&= \mathbb{E}\left(\ln e^{-T/\tau_{ca}}\right), \\
&= \mathbb{E}\left(\frac{-T}{\tau_{ca}}\right), \\
&= \frac{-1}{\tau_{ca}}\mathbb{E}\left(T\right),
\end{aligned}$$

because T is an exponentially distributed random variables with rate parameter λ, its mean is equal to reciprocal of its rate parameter , i.e. λ^{-1}. Therefore, we have

$$\mathbb{E}\left(\ln|A|\right) = \frac{-1}{\tau_{ca}\lambda},$$

where $\lambda > 0$ and $\tau_{ca} > 0$. Hence

$$\mathbb{E}\left(\ln|A|\right) < 0.$$

Earlier in this section, we assumed that the change in calcium concentration Δ is a random variable. For a reason that we will justify later in this section, we consider that this random variable follows a gamma distribution with rate and shape parameter 1, denoted as $\Delta \sim Gamma(\alpha = 1, \lambda = 1)$ which is simpler to call Δ is an exponentially distributed random variable with rate parameter $\lambda = 1$.

The second condition is that $\mathbb{E}\left(\ln|\Delta|\right) < \infty$. In order to show that the second condition holds, we start by stating a simple theorem concerning the logarithmic expectation of gamma random variable.

Theorem. *The expected value of the natural logarithm of a gamma random variable X with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$ is formulated as*

$$\mathbb{E}\left(\ln X\right) = \psi\left(\alpha\right) - \ln(\lambda), \tag{4.2.5}$$

where $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digamma function.

Proof. The probability density function for the gamma random variable is

$$\begin{cases} f(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{for } x > 0 \text{ and } \alpha, \lambda > 0, \\ 0 & x \leq 0 \end{cases} \tag{4.2.6}$$

where $\Gamma(\alpha)$ is the gamma function.

The expectation of natural logarithm of gamma random variable X is written as

$$\mathbb{E}\left(\ln(X)\right) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} e^{-\lambda x} \ln x dx.$$

Let $z = \lambda x$. Then,

$$\begin{aligned}
\mathbb{E}\left(\ln(X)\right) &= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} (\frac{z}{\lambda})^{\alpha-1} e^{-\lambda z/\lambda} \ln \frac{z}{\lambda} \frac{dz}{\lambda}, \\
&= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} z^{\alpha-1} e^{-z} \ln \frac{z}{\lambda} dz, \\
&= \ln(\lambda^{-1}) \underbrace{\int_0^{\infty} \frac{1}{\Gamma(\alpha)} z^{\alpha-1} e^{-z} dz}_{1} + \frac{1}{\Gamma(\alpha)} \underbrace{\int_0^{\infty} z^{\alpha-1} e^{-z} \ln z dz}_{\Gamma(\alpha)\psi(\alpha)}, \\
&= -\ln(\lambda) + \psi(\alpha).
\end{aligned}$$

□

Therefore, given $\Delta \sim Gamma(1,1)$ by eqn. (4.2.5), $\mathbb{E}\left(\ln|\Delta|\right)$ is

$$\begin{aligned}
\mathbb{E}\left(\ln|\Delta|\right) &= \psi(1) - \ln(1), \\
&= \psi(1), \\
&= \frac{\Gamma'(1)}{\Gamma(1)}, \\
&= -\gamma,
\end{aligned}$$

where $\gamma \approx 0.577215\cdots$ is the Euler-Mascheroni constant.

Thus, by the fundamental theoretical result in [34], we infer from previous subsection that $C_n \xrightarrow{d} C$, then C satisfies the distributional identity

$$C \overset{d}{=} AC + \Delta, \quad C \text{ and } (A, \Delta) \text{ independent.}$$

Following [70] let $X := AC$, then we can write

$$X \stackrel{d}{=} A\,(X + \Delta)\,, \tag{4.2.7}$$

and hence

$$C \stackrel{d}{=} X + \Delta.$$

Iterating eqn. (4.2.7) results in

$$X \stackrel{d}{=} \sum_{n=1}^{\infty} A_1 A_2 . \cdots . A_n \Delta_n$$

where $A_n = e^{-T_n/\tau_{ca}}$.

The theorem below indicates that if T_n is exponentially distributed with rate parameter λ, then random variable A_n has Beta distribution with shape parameters ($\lambda\tau_{ca}$,1) and is denoted by $A_n \sim Beta\,(\lambda\tau_{ca}, 1)$.

Note that probability density function for a Beta distributed random variable X with shape parameters α and β is given by

$$f\,(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad \text{for } \; 0 < x < 1 \text{ and } \alpha, \beta > 0$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Theorem. *If X is an exponentially distributed random variable with rate parameter λ, then for every $c > 0$, $Y = e^{-X/c}$ follows the beta distribution with shape parameters λc and 1.*

Proof. The transformation $Y = g(X) = e^{-X/c}$ is 1-1 transformation from $X = \mathbb{R}^+$ to $y = [0, 1]$. The inverse of the transformation is $X = g^{-1}(Y) = -c \ln Y$, and the associated Jacobian

is $\frac{dX}{dY} = \frac{-c}{Y}$. By the transformation theorem, the probability density function of Y is

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(X)) \left|\frac{dx}{dy}\right|, \\ &= \lambda e^{-\lambda c \ln y} \frac{c}{y}, \\ &= c\lambda y^{c\lambda - 1}, \\ &= \frac{1}{B(c\lambda, 1)} y^{c\lambda - 1}, \end{aligned}$$

where $B(c\lambda, 1) = \frac{\Gamma(c\lambda)\Gamma(1)}{\Gamma(c\lambda+1)}$ and Γ is gamma function. Note that for any $\alpha > 0$, the gamma function satisfies the recursive property $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$. Thus here we have

$$B(c\lambda, 1) = \frac{\cancel{\Gamma(c\lambda)}\overbrace{\Gamma(1)}^{1}}{c\lambda\cancel{\Gamma(c\lambda)}} = \frac{1}{c}$$

Thus $Y \sim Beta(c\lambda, 1)$. □

Therefore, $A = e^{-T/\tau_{ca}}$ follows Beta distribution with shape parameters $\lambda\tau_{ca}$ and 1.
By applying a beta-gamma algebra identities ([20]) which state that for some $a, b > 0$, we have

$$Beta\,(a, b) \odot Gamma\,(a + b, 1) \overset{d}{=} Gamma\,(a, 1)\,.$$

Note that the notation "$X \sim f(x)$" stands for "the variable X has probability distribution f. "If $X_i \sim f_i, i = 1, 2$, are independent, then the distribution of $X_1 X_2$ is denoted $f_1 \odot f_2$.
Also,

$$Beta\,(a, b) \odot (Gamma\,(a, 1) + Gamma\,(b, 1)) \overset{d}{=} Gamma\,(a, 1)\,. \tag{4.2.8}$$

Applying eqn. (4.2.8) in eqn. (4.2.7), and considering that $A_n \sim Beta\,(\lambda\tau_{ca}, 1)$, then

$$X \sim Gamma\,(\lambda\tau_{ca}, 1)\,,$$
$$\Delta \sim Gamma\,(1, 1)\,.$$

Note that $C \stackrel{d}{=} X + \Delta$ implies

$$C \stackrel{d}{=} Gamma(\lambda\tau_{ca}, 1) + Gamma(1, 1).$$

Finally

$$C \sim Gamma(\lambda\tau_{ca} + 1, 1).$$

4.2.3 Assessing the fit of the calcium concentration distribution with the Kolmogorov-Smirnov (k-s) test

Here we use one sample *Kolmogorov-Smirnov* test to assess evidence supporting the calcium concentration follows a gamma distribution. The application of this test involves a simple calculation comparing the empirical cumulative distribution function of calcium concentration data, F_{obs}, with the cumulative distribution function associated with the null hypothesis, F_{exp}.

Let $c_1, c_2, \cdots, c_n$ be calcium observations on continuous i.i.d. r.vs $C_1, C_2, \cdots, C_n$ with common cumulative density function F. We want to test the hypothesis

$$H_0 : F(c) = F_{exp}(c) \text{ for all } c,$$

where F_{exp} is gamma CDF.

The Kolmogorov-Smirnov test statistic D is defined by

$$D = \underset{c \in \mathcal{R}}{sup} |F_{obs}(c) - F_{exp}(c)|,$$

where F_{obs} is an empirical cumulative distribution defined as

$$F_{obs}(c) = \frac{\#(i : c_i \leq c)}{n}$$

Note that the p-value returned by the k-s test has the same interpretation as other p-values. Table 5.6.4 displays the results of Kolmogorov-Smirnov for calcium data at different mean firing rates. 20000 simulations were performed using stochastic map to generate calcium concentration data for mean firing rates 0.5, 3, 8, 10, 20, 100, 5000, 8000. It is observed that with such large p-values, there is not enough evidence to reject the null hypothesis at 5% for all mean firing rates. Therefore, it is concluded that the simulated calcium concentration are gamma distributed. In addition, gamma distribution matches simulation results well over a range of parameter values. Simulation and analytic results in Figure 4.1 indicates the good of fit of the gamma distribution to the calcium concentration.

Table 4.2.1: Kolmogorov-Smirnov test gamma distributed on Calcium data

ν	Statistics (D)	p-value
$\nu = 0.5$	0.00805	0.5360
$\nu = 3$	0.006	0.8643
$\nu = 8$	0.0066	0.7764
$\nu = 10$	0.00945	0.3337
$\nu = 20$	0.0034	0.9998
$\nu = 100$	0.0071	0.6945
$\nu = 5000$	0.0103	0.2392
$\nu = 8000$	0.0252	0.1097

D is the Kolmogorov-Smirnov Statistic

The hypothesis to be tested was formulated as;

H_0: Calcium concentrations are gamma distributed vs

H_1: Calcium concentrations are not gamma distributed.

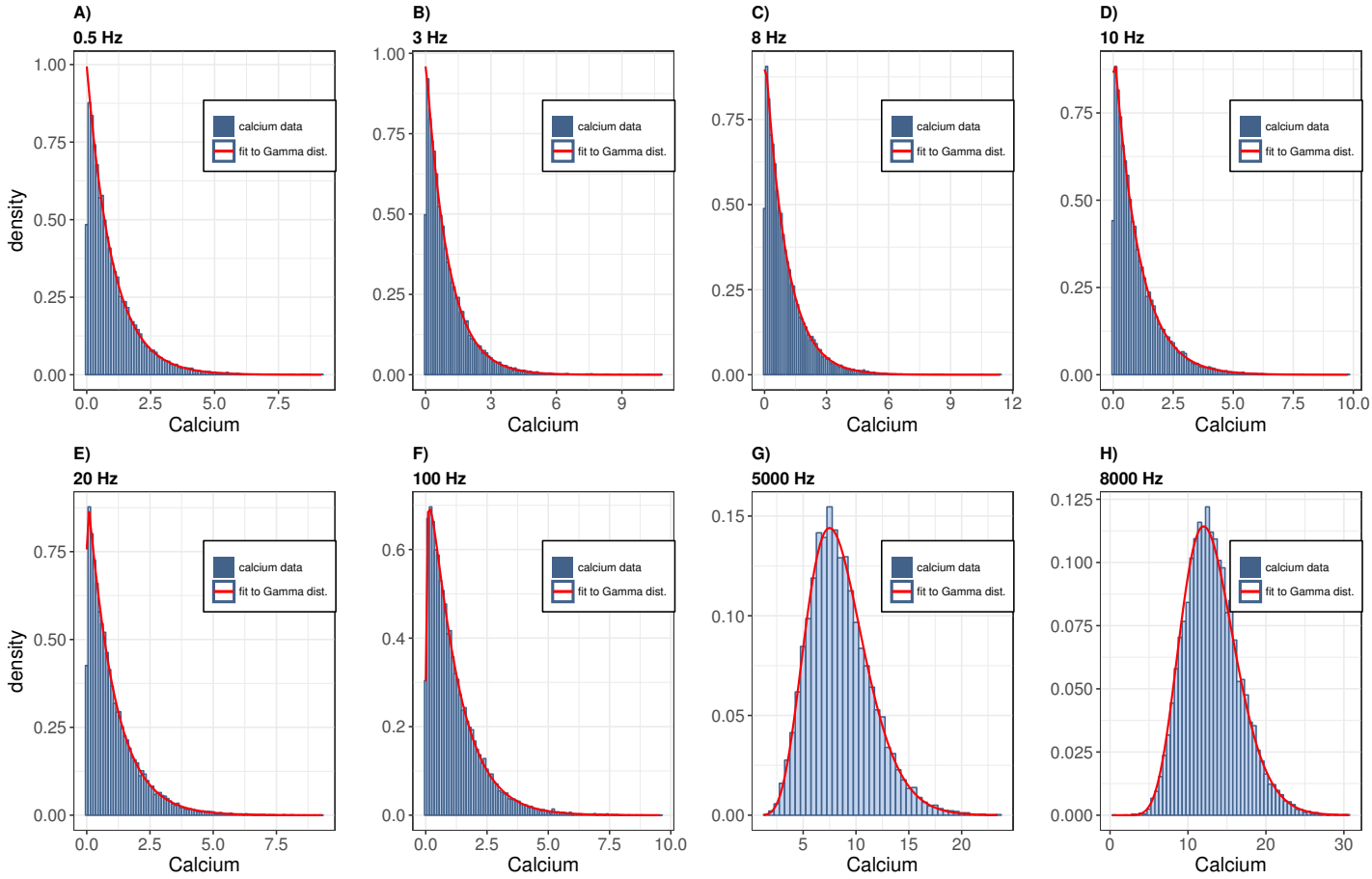


Figure 4.1: Histograms of calcium concentration data for mean firing rates 0.5, 3, 8, 10, 20, 100, 5000 and 8000 Hz, together with fitted gamma pdfs. Plots show adherence to a linear relationship between the simulated and theoretical quantiles, confirming our analytic results.

4.2.4 Quantile Gamma Graph Plot

We compare this result with the distribution obtained by numerical simulation of the recurrence relation for C by creating qq-plots. Figure 4.2 displays quantile plots for the map with input mean firing rates 0.5, 3, 8, 10, 20, 100, 5000, 8000 Hz, with the theoretical quantiles based upon the gamma distribution. This graphical display shows whether the simulated data can reasonably be described by a gamma distribution. These plots indicate the simulated data aligns with gamma distributed random variables in a straight line, indicating that the calcium

concentration have a gamma distribution. It is clear that when there is confidence that data are sampled from a family of distributions described by some parameters it is possible to use that information to obtain more data-efficient estimators such as mean.

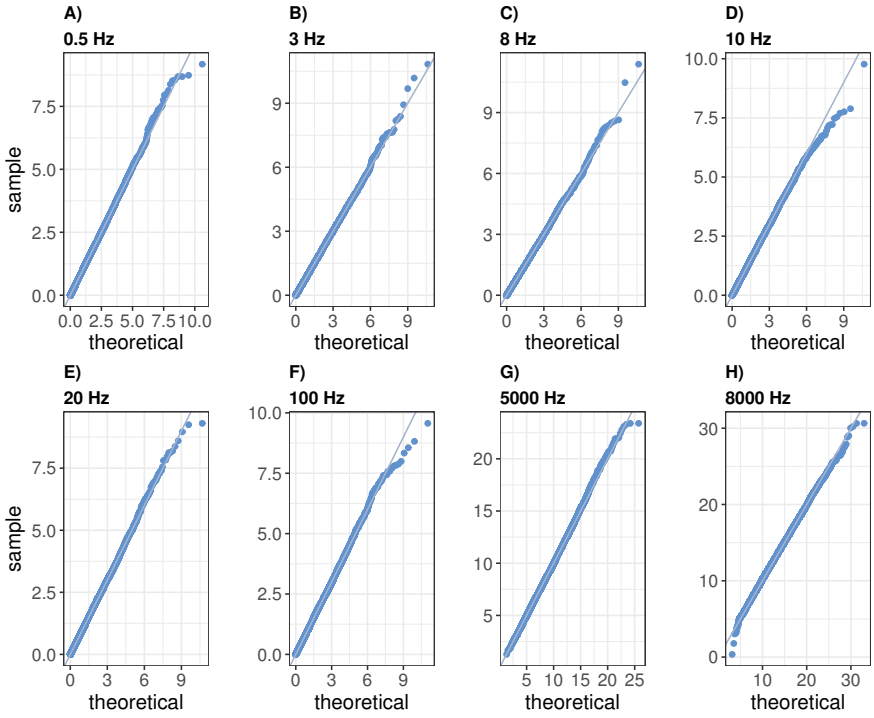


Figure 4.2: Gamma qq-plot of calcium concentration for 0.5, 3, 8, 10, 20, 100, 5000 and 8000 Hz mean firing rates.

4.3 Distribution of P_r for small τ_{ca} and Large Interspike Interval T

We conclude from the previous subsection that the calcium concentration does have a closed form distribution, and indeed a gamma distribution when the data is simulated according to the model in hand. However, this is not the case for the variable R due to the complexity of the map, and so a closed form for the distribution of $P_r = PrR$ is not possible. However, we can understand it partially by considering the mechanisms involved, as done in the preceding section. We can also use some information from the deterministic map. The map has a single attracting fixed point, and the collapse to this fixed point from physiological initial conditions is very rapid [66]. The value of the fixed point depends on the firing rate, with a smaller value for larger firing rate in general. In Figure 4.3 we plot the expression for the fixed point of the deterministic map vs. rate, along with the mean of the distribution of P_r for varying frequencies. The values decrease with increasing frequency, as expected, and are remarkably close.

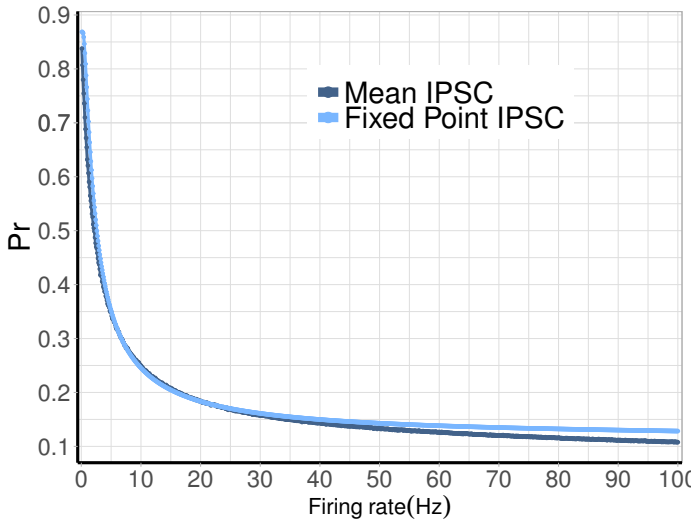


Figure 4.3: Fixed point values and mean of P_r for the deterministic map as a function of mean firing rates.

This motivates the idea that if the P_r value was directly determined by the fixed point value for the ISI value preceding it, we would be able to approximate distribution of P_r through the fixed point. We examine this when the calcium decay time (τ_{ca}) is notably smaller than the inter-spike interval (T). With this approximation C, Pr and R have time in between pulses to decay to their steady state value before another pulse. This means that the fixed point value for a rate is given by $1/T$ where T is the preceding interspike interval is more likely to give a good estimate of the actual value of $P_r = PrR$.

It was shown in [66] that the fixed point of the map which represents the peak IPSC over long term is

$$\begin{aligned}\overline{C} &= \frac{\Delta}{1-e^{-T/\tau_{ca}}},\\ \overline{Pr} &= \frac{P_{max}\overline{C}^4}{\overline{C}^4+K^4},\\ \overline{R} &= \frac{1-\gamma(\overline{C})}{1-\gamma(\overline{C})(1-\overline{Pr})},\end{aligned}$$

where

$$\gamma(\overline{C}) = \left(\frac{\overline{C}e^{-T}+K_r}{\overline{C}+K_r}\right)^{\Delta k} e^{-k_{min}T}.$$

As T increases $\overline{C} \to \Delta$ and hence $\overline{Pr} \to P_{max}$.

Therefore, the fixed point $\overline{R}$ is then

$$\bar{R} = \frac{1-e^{-k_{min}T}}{1-(1-P_{max})e^{-k_{min}T}}.$$

From this we can compute the probability density function of $\overline{R}$, given an exponential distribution for the variable T. For the ease of notation, let assume that $X = \overline{R}$ is a random

variable and is defined as follows

$$X = \frac{1 - e^{-k_{min}T}}{1 - (1 - P_{max})e^{-k_{min}T}}.$$

Also, we assumed that random variable T have exponential distribution with probability density function

$$f_T(t) = \lambda e^{-\lambda t} \quad t > 0.$$

We can compute an analytical expression for probability density function (PDF) of fixed point $\overline{R}$ using the distribution for T.

The transformation $X = g(T) = \frac{1-e^{-k_{min}T}}{1-(1-P_{max})e^{-k_{min}T}}$ is a 1-1 transformation from $\mathcal{T} = \{t|\ t > 0\}$ to $\mathcal{X} = \{x|\ 0 < x < 1\}$ with inverse $T = g^{-1}(X) = \frac{1}{k_{min}} \log\left(\frac{1-(1-P_{max})x}{1-x}\right)$ and Jacobian

$$\frac{dT}{dX} = \frac{1 - (1 - P_{max})}{k_{min}(1 - x)(1 - (1 - P_{max})x)}$$

By the rule for functions of random variables, the probability density function of X is

$$\begin{aligned} f_X(x) &= f_T\left(g^{-1}(x)\right)\left|\frac{dt}{dx}\right| \\ &= \frac{\lambda\left(1 - (1 - P_{max})\right)}{k_{min}}\left(1 - x\right)^{-(1-\lambda/k_{min})}\left(1 - (1 - P_{max})x\right)^{-(1+\lambda/k_{min})} \end{aligned}$$

Thus, an analytic expression for its probability density function (PDF) exists and is given by

$$f\left(x|\lambda, c, k_{min}\right) = \frac{\lambda(1-c)}{k_{min}}(1-x)^{-(1-\lambda/k_{min})}(1-cx)^{-(1+\lambda/k_{min})}, \tag{4.3.1}$$

where $c = 1 - P_{max}$, $\lambda > 0$ is the rate and $k_{min} > 0$ is the baseline recovery rate. The distribution is supported on the interval $[0, 1]$.

From the expression 4.3.1, the expected value of random variable $X = \overline{R}$ is given by

$$E(X) = (1-c)\lambda \left(\frac{1}{\lambda(1-c)} - \frac{{}_2F_1\left(1, \frac{k_{min}+\lambda}{k_{min}}; 2 + \frac{\lambda}{k_{min}}; c\right)}{k_{min}+\lambda} \right),$$

where ${}_2F_1\left(1, \frac{k_{min}+\lambda}{k_{min}}; 2 + \frac{\lambda}{k_{min}}; c\right)$ is the hypergeometric function.

Similarly, we can compute the analytical expression of the probability density function of fixed point $Y = \overline{PrR}$. We will refer to this in what follows as the *stochastic fixed point*. Hence, the probability density function for the stochastic fixed point is

$$f(y|\lambda, c, k_{min}) = \frac{\lambda P_{max}(1-c)}{k_{min}} (P_{max} - y)^{-(1-\lambda/k_{min})} (P_{max} - cy)^{-(1+\lambda/k_{min})}.$$

This distribution is supported on the interval $[0, P_{max}]$. Figure 4.4 shows this expression for different mean input inter-spike interval, in milliseconds.

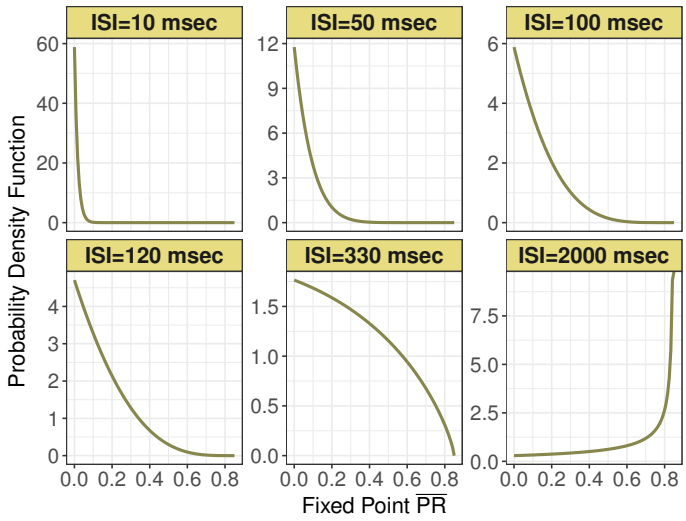


Figure 4.4: PDF of the stochastic fixed point $\overline{PrR}$ for varying interspike intervals of 10, 50, 100 , 120, 330, 2000 in millisecond. Parameters $k_{min} = 0.0013$ and $P_{max} = 0.85$, are from the control set.

Figure 4.5 are histograms of P_r values obtained from the map with very small τ_{ca}, and

other parameters from the control set, as in Figure 4.4. The similarity between the two is evident. Apparently this approximation captures not only the mean value of the numerical distribution, but also the shape of the distribution and how it changes with varying input spike train rate.

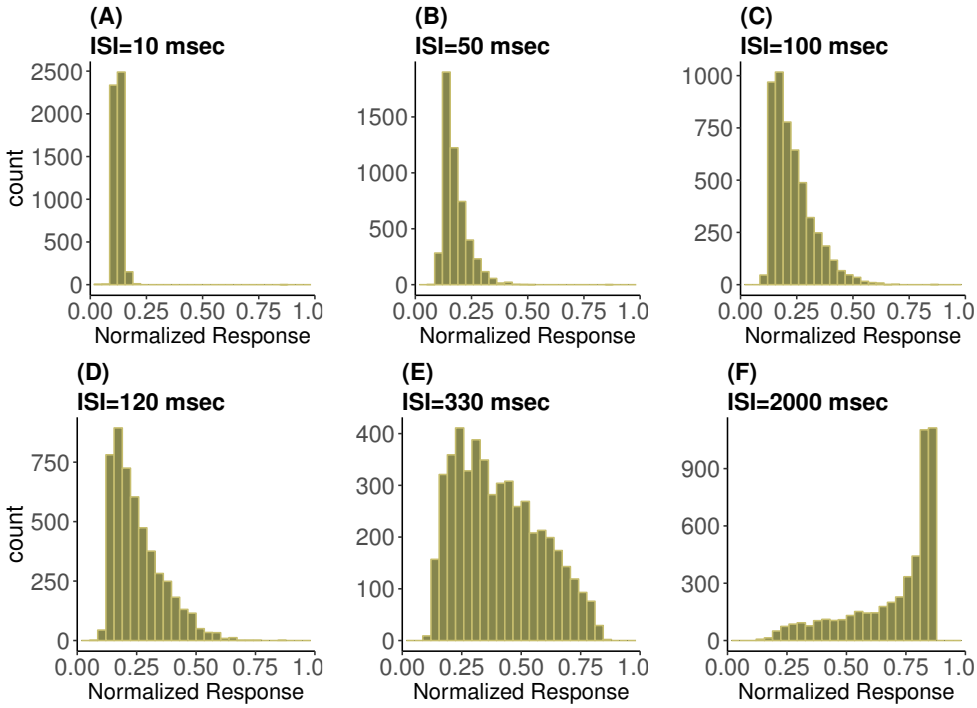


Figure 4.5: Frequency Distribution of P_r for varying mean input ISI A) 10 B) 50 C) 100 D) 120 E) 330 and F) 2000 in milliseconds, when interspike interval T is significantly larger that the calcium decay time τ_{ca}

Figure 4.6 compares these two distributions via quantile-quantile plots, for mean ISI of 10, 50, 100, 120, 330, and 2000 milliseconds. In every QQ-plot the quantiles of all $\overline{PR}$ are plotted against the quantiles of all P_r values. If the values of the two different data sets have the same distribution, the points in the plot should form a straight line. From these plots it is clear that when the mean ISI is significantly larger than calcium decay time, the distribution of the stochastic fixed point is similar to that of P_r. However, for smaller mean ISI (10 msec) the approximation becomes less exact, so the similarity between two distributions decreases.

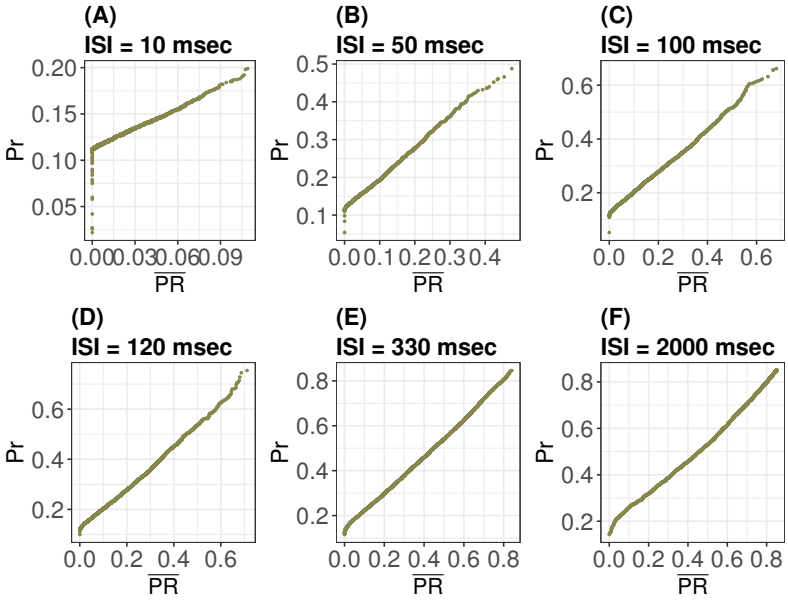


Figure 4.6: Quantile plot of two data sets obtained from Pr and $\overline{PR}$.

4.4 Discussion

Creating closed form expressions for P_r distributions in not possible as the map is too complex. However, the stochastic recurrence relation for the calcium concentration alone is simple enough to be analyzed and we discovered that it follows a gamma distribution with shape and scale parameters $\lambda\tau_{ca}+1$ and 1, respectively. For the P_r distribution we had to rely on an approximation motivated by numerical results. We found the mean of the distribution followed the frequency in the same way that the fixed point of the map did. The collapse to the fixed point is very rapid which justifies our assumption that the P_r values are determined by the fixed point value associated with the instantaneous rate of the preceding inter spike interval. Then the formula for the fixed point as a function of frequency could be used to generate a distribution, using the exponential distribution of the T's. The results confirm the validity of this approximation and most importantly, the "sloshing" of the distribution between zero and P_{max} as the frequency that is decreased through the physiological range is captured by this form.

Chapter 5

Data Driven Models of Synaptic Plasticity

5.1 Motivation

It is important to recognize that neurons are computational devices and convey the results of their computations in the form of the output spike trains. We often observe the output spikes, however the input is almost always unknown and thus, it is very difficult to determine the functional form of these computations. In previous chapters we used information theoretic basics to determine, via the estimation of entropy and mutual information, the amount of information presynaptic neuron can transmit, given Poisson stimulated spikes. These measures mainly quantify randomness and provide limited understanding about the structure of output spike train or the amount of computation required to produce this structure. In this chapter, we introduce the most compact and sufficient description of a process capable of statistically reproducing the observed large "1" and small "0" postsynaptic response output. We will do this through identifying the minimal hidden Markov Model that generates binary postsynaptic

responses and can statistically predict the future of the response without loss of information.

5.2 Theoretical Foundations

According to Shalizi and Crutchfield [57, 16] we define some important concepts of computational mechanics.

5.2.1 Causal States

Consider a stochastic process give by consecutive discrete random variables, $\cdots, S_{-2}S_{-1}S_0S_1S_2\cdots$, where each S_i may take a symbol s_i drawn from a finite countable set $\mathcal{A}$ of size k. At any time t, we can divide the sequence into a past (history) $\overleftarrow{S}_t$ and a future $\overrightarrow{S}_t$. If the process is conditionally stationary; i.e. for all possible future events $\mathbb{A}$, $P(\overrightarrow{S}_t \in \mathbb{A} | \overleftarrow{S}_t = \overleftarrow{s})$ does not depend on t, then we can drop the subscript and use $\overleftarrow{S}$ and $\overrightarrow{S}$ instead. Note that $\overleftarrow{S}^L$ denotes the last L symbols of the history, while $\overrightarrow{S}^L$ refers to the first L symbols of the future. Two histories $\overleftarrow{s}$ and $\overleftarrow{s'}$ are equivalent if $P(\overrightarrow{S} = \overrightarrow{s} | \overleftarrow{S} = \overleftarrow{s}) = P(\overrightarrow{S} = \overrightarrow{s} | \overleftarrow{S'} = \overleftarrow{s'})$, i.e., when they share the same distribution for the future. We can define a function that maps histories to their equivalence classes:

$$\epsilon(\overleftarrow{s}) = \{\overleftarrow{s'} : P(\overrightarrow{S} = \overrightarrow{s} | \overleftarrow{S} = \overleftarrow{s'}) = P(\overrightarrow{S} = \overrightarrow{s} | \overleftarrow{S} = \overleftarrow{s})\}$$

The range of the function ϵ, groups of histories sharing future distributions, are named *causal states* of the process. We denote the i^{th} causal state as σ_i and the set of all causal states as $\boldsymbol{\mathcal{S}}$; the corresponding random variable is denoted $\mathcal{S}$ and its realization σ.

5.2.2 Causal States Transitions

At any point in time, the current causal state and the next symbol in the sequence determine the next causal state. Therefore, we can define state-to-state transitions between causal states and the probabilities of these transitions. The probability of moving from state σ_i to state σ_j on symbol s is

$$T_{ij}(s) = P(\overrightarrow{S}^1 = s, \mathcal{S} = \sigma_j | \mathcal{S} = \sigma_i)$$

Note that

$$\sum_{s \in \mathcal{A}} \sum_{\sigma_j \in \boldsymbol{\mathcal{S}}} T_{ij}(s) = \sum_{s \in \mathcal{A}} P(\overrightarrow{S}^1 = s | \mathcal{S} = \sigma_i) = 1.$$

We can now define a machine that combines the set of causal states and transitions that represent the process. This machine is called ϵ-machine [17].

5.2.3 Properties of Causal States

Causal states have many important properties that make them a representation of a process. Here we state main properties of causal state See [57] for proofs.

- Causal states are minimal sufficient statistics for predicting the process's future.
- Given an initial Causal State and the next symbol from the original process, only certain successor causal states are possible.
- The causal states are homogeneous over future events: That is, all histories belonging to a single causal state σ have the same conditional distribution of future events.
- Strict Homogeneity of Causal States: A process's causal states are the largest subsets of inputs that are all strictly homogeneous with respect to the future event.
- Each causal state has a unique morph, i.e., no two causal states have the same conditional

distribution of futures.

5.2.4 ϵ-machine Reconstruction

Any procedure that produces ϵ-machine which represents a given process is ϵ-machine reconstruction. There are plenty of algorithms that reconstruct ϵ-machine based on the data sequences. In section 5.3 we use an algorithm that estimate an ϵ-machine from samples of a process, while respect the necessary properties of causal states outlined above.

5.2.5 Statistical complexity

From the constructed ϵ-machine, $P(\sigma_i)$, the probability of finding a system in the causal state i after the machine has been running infinitely long is given by the left eigenvector of transition matrix T with eigenvalue 1, normalized in probability. That is

$$P(\sigma_i) = \sum_{i=1}^{\|\mathcal{S}\|} P(\sigma_i) T_{ij},$$

where $\|.\|$ represents the carnality of a set. Thus the Statistical Complexity is defined as

$$C_\mu = -\sum_i P(\sigma_i) \log_2 P(\sigma_i).$$

C_μ measures the minimum amount of historical information needed to make optimal prediction. In other words, statistical complexity is able to quantify the degree of physical structure present in a time series.

5.3 Causal-State Splitting Reconstruction (CSSR)

In this section we describe an algorithm introduced by Shalizi [59] that infers the causal states of a process and the transition probabilities between causal states from sequential data. It builds the minimal set of hidden Markovian states that is capable of producing the behavior presented in the data. This algorithm makes no prior assumption about the process's causal structure (the number of hidden states and their transition structure), but infers this from the observed data. The input and output of the algorithm are:

- **Input:** Length (N) of the data sequence, the measurement alphabet size (k), maximum history length (L_{max}), and the significant level for the hypothesis test (α).
- **Output:** A set of estimated causal states $\hat{\sigma}_i$ and the transition matrix T.

5.3.1 The Algorithm

CSSR algorithm starts by assuming the process is an independent and identically distributed (IID) sequence and all the histories belong to the single state, and then successively tests whether longer and longer suffixes result in the conditional distribution for the next observation which differ significantly from the state to which they currently belong and thus adds new states. The Causal State Machine (CSM) is divided into three steps: Initialize, Homogenize and Determinize.

1. **Initialize:** Initially we assume that the process is an IID sequence. ($L = 0$, $\hat{\boldsymbol{\mathcal{S}}} = \{\hat{\sigma_0}\}$, where $\hat{\sigma_0} = *\lambda$); i.e., $\hat{\sigma_0}$ contains only null sequence λ. Note that $*\lambda$ is a suffix of any history, so initially all the histories are mapped to this single state. We can calculate

the morph of this state as

$$P(\overrightarrow{S}^1 = a|\hat{\sigma}_0) = P(\overrightarrow{S}^1 = a). \tag{5.3.1}$$

The initial model can be thought of as tossing a fair coin long enough and try to predict the outcome in the coin flip. It is clear that we do not need to keep track of outcome because the flips are independent. Thus, having knowledge of previous tosses does not reduce the uncertainty about the next toss. Hence, in this case no memory is needed to optimally predict the next observation. As a consequence, the statistical complexity vanishes ($C_\mu = 0$).

2. **Homogenize:** We generate states whose histories lead to the same morph; i.e., states whose members of histories have no significant difference in their individual morphs. This procedure is as follows.

   - For each $\hat{\sigma} \in \hat{\boldsymbol{\mathcal{S}}}$, compute state morph.

   $$\hat{P}(\overrightarrow{S}^1|\hat{\mathcal{S}} = \hat{\sigma})$$

   Note that since the conditional distribution over the futures can be seen as "the shape of the future" , we call this the state's "morph".

   (a) For $L = 0$ we use eqn. 5.3.1.

   (b) For each sequence $\overleftarrow{s}^L \in \hat{\sigma}$, estimate the morph of the history $\overleftarrow{s}^L$. This can be calculated as follows,

   $$\hat{P}(\overrightarrow{S^1} = a|\overleftarrow{S}^L = \overleftarrow{s}^L) = \frac{\nu(\overleftarrow{S}^L = \overleftarrow{s}^L, \overrightarrow{S}^1 = a)}{\nu(\overleftarrow{S}^L = \overleftarrow{s}^L)} \tag{5.3.2}$$

   (c) The morph of the state $\hat{\sigma}$ is the weighted average of the morphs of its histories

$\overleftarrow{s}^L \in \hat{\sigma}$, with weights proportional to $\nu(\overleftarrow{S}^L = \overleftarrow{s}^L)$

$$\hat{P}(\overrightarrow{S}^1 = a|\hat{\mathcal{S}} = \hat{\sigma}) = \frac{1}{z} \sum_{\overleftarrow{s}^L \in \hat{\sigma}} \nu(\overleftarrow{S}^L = \overleftarrow{s}^L) \hat{P}(\overrightarrow{S}^1 = a|\overleftarrow{S}^L = \overleftarrow{s}^L),$$

where $z = \sum_{\overleftarrow{s}^L \in \hat{\sigma}} \nu(\overleftarrow{S}^L = \overleftarrow{s}^L)$ is the number of occurrence in $s^N, s \in \mathcal{A}$ of suffixes in $\hat{\sigma}$

- For each $\hat{\sigma} \in \hat{\mathcal{S}}$, test the null hypothesis (similarity in histories' morph). For each L history $\overleftarrow{s}^L \in \hat{\sigma}$ and each $a \in \mathcal{A}$, generate the suffix $a\overleftarrow{s}^L$ of length $L+1$.
  (a) Estimate the morph of $a\overleftarrow{s}^L$ using eqn. 5.3.2.
  (b) If the morphs of $a\overleftarrow{s}^L$ and $\hat{\sigma}$ do not differ according to the significance test, add $a\overleftarrow{s}^L$ to $\hat{\sigma}$.
  (c) If they do differ, test whether there are any states in $\hat{\boldsymbol{\mathcal{S}}}$ whose morphs do not differ significantly from that of $a\overleftarrow{s}^L$. If so, add $a\overleftarrow{s}^L$ to the state whose morph its morph matches most closely, as measured by the score of the significance test.
  (d) If the morph of $a\overleftarrow{s}^L$ is significantly different from the morphs of all existing states, then create a new state and add $a\overleftarrow{s}^L$ to it.
  (e) Recalculate the morphs of states from which sequences have been added or deleted.
- Increment L by one.
- Repeat above steps until reaching the maximum history length L_{max}.

3. **Determinize**: Split the states until they have deterministic transitions. To do this, in each state we compute transitions for each suffix. If two suffixes in one state transit to different state for the same symbol, we split them into two different states.

5.3.2 Parameter selection for CSSR Algorithm

CSSR has two user-specified parameters. The significant level α which is assigned when we use Kolmogorov-Smirnov (KS) test to decide whether the estimated morphs of histories' subsequences are significantly differ from all other state's morph. In case of significant difference, new states is formed for these subsequences. By assigning different values to significant level α, we control the risk of seeing structure that is not there and the states are merely created due to sampling error, rather than the actual differences between their morphs. Some common choices of α that work well are 0.001, 0.01, 0.1 and 0.05. If α is small, we need larger statistics to reject the null hypothesis and split the the states.

Also, the CSSR Algorithm depends crucially on another user-set parameter L_{max} which is the maximum subsequence length considered when inferring the model structure (machine). Setting L_{max} too large results in data shortage for long strings and algorithm tends to produce too many states and hence results become unreliable. On the other hand, if L is too small, the algorithm won't be able to capture all statistical dependencies in the data and hence the state structure of the inferred machine is not valid. Shalizi [59] showed that there is a lower bound for the acceptable value of L_{max}; i.e., it must be large enough that every state contains at least one suffix of that length. However, an optimal choice of L_{max} is not straight forward. One approach which is used in this work is to determine the longest history length according to the relationship derived from Hanson [28]. Hanson showed that, for a given fixed finite amount of data N, and fixed significance level α, we choose the maximum length of subsequence L such that

$$\sqrt{\frac{|\mathcal{A}|^{L_{max}}}{N - L_{max}}} = \alpha. \tag{5.3.3}$$

5.4 Study of CSSR to Capture the Structure of PSRs

In this section, we study dynamical aspects of short-term synaptic plasticity using the CSSR algorithm. This algorithm describes synaptic dynamics as a hidden Markov process and illustrates how to infer a model of the hidden process that generated the observed behavior. The ultimate goal is a categorization of the types of processes a synapse can create, and an assignment of those to different synapse types under varying conditions.

5.4.1 Analyzing the map

A synapse can be classified as being "depressing", "facilitating" or "mixed" depending on its response to stimulation at a relevant frequency. The model here [66] is built so that, depending on the parameters, facilitation, depression, and mixture of both is possible. For instance, by varying the parameters we can create a "mock" facilitating synapse, where the size of response increases with increasing input frequency, or a mixed synapse, where the response is decreased for low and high frequency, but increases for moderate values of the frequency. Table 5.4.1 shows the parameter values for each synapse type we considered.

Table 5.4.1: Parameter values for "mock" synapses

parameter	facilitating	mixed
K	4.0	1.0
k_{min}	0.002 1/msec	0.002 1/msec
k_{max}	6.0 1/msec	6.0 1/msec
K_r	0.1	0.1
τ_{Ca}	30 msec	30 msec
P_{max}	0.6	0.6

Figure 5.1 shows the fixed point for different types of synapse dynamics as a function of input mean firing rates. It can be seen that the depressing synapse fixed point decreases quickly from P_{max} over $0 - 10$ Hz, and then decays slowly to almost zero for higher mean firing rates. The facilitating synapse fixed point however, increases over the physiological range, but decreases very slowly for larger values of the frequency. The mixed synapse fixed point starts at a base value of 0.2, increases to a local maximum near 25 Hz and decays thereafter. The competition between increasing release probability and decreasing fraction of release ready vesicles creates the local maximum in the mixed synapse, see Figure 5.2.

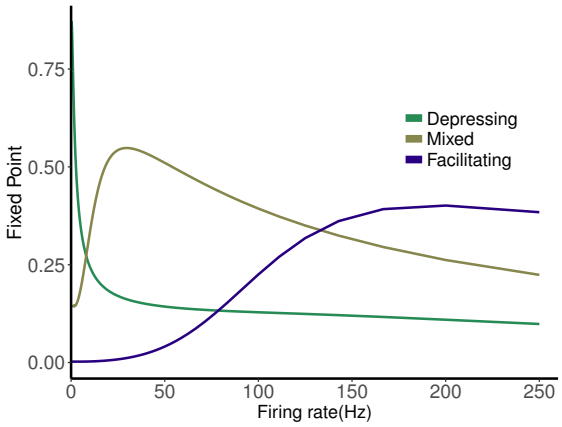


Figure 5.1: Fixed point values of normalized postsynaptic response for three synapse models of “depressing”, “mixed”, and “facilitating ” stimulated by Poisson spike trains with mean firing rates ranging from 0.1 to 250.

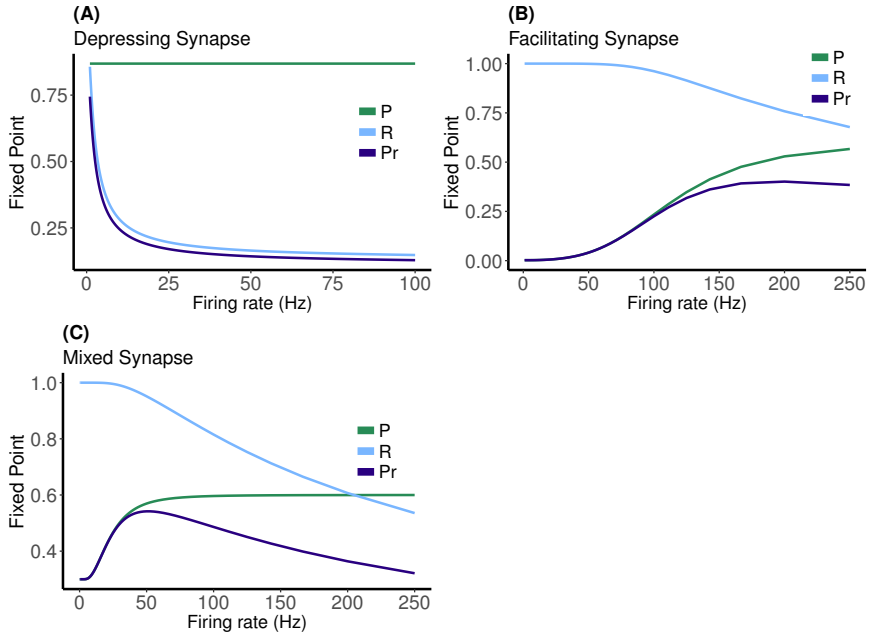


Figure 5.2: Fixed point values for release probability P, fraction of readily releasable pool R and normalized postsynaptic response Pr for varying mean firing rates ranges from 0.1 to 100 Hz for **(A)** depressing synapse and from 0.1 to 250 Hz for **(B)** facilitating and **(C)** mixed synapse.

5.4.2 Method

Here we consider output responses of different types of synapses as a stochastic binary time series where "1" corresponds to large response and "0" to relatively small response. Our goal is to find a minimal representation of the computational structure present in this time series. In other words, the structure present in the output responses can be described by the CSSR algorithm. Note that the causal states created are optimal predictors of the output response's future, using information from the response's history. Each causal state is defined by grouping histories of past response activity which are statistically equivalent in terms of predicting the future response. The CSM is represented by a directed graph, with nodes indicating the process's hidden states and edges the transitions between states. Each edge is labeled with symbol emitted during the transition ("1" for large response, "0" for small response), and the

probability of state-to-state transition.

Partition

One of the CSM assumptions is that the observed process is formed by sequence of discrete symbols, so it is not designed for processes with continuous variables. Thus, in order to reconstruct CSMs, the output responses need to be partitioned into a sequence of 1's and 0's, which requires the consideration of a threshold value. There are many threshold values approaches are available such as half P_{max} cut off, median, mean or fixed point value for the map. However, since we want to extract the maximum structure present in the observed data, and statistical complexity quantifies the degree of this structure, we use the threshold within $[0, 1]$ which provides maximum statistical complexity in the resulting machine.

Reconstructed Causal State Machines

After partitioning the output responses into a sequence of 0's and 1's, we apply the CSSR algorithm [58], using Causal State Modeller Toolbox in Matlab [33] to built machines for the three different types of synapse. We use $N = 10^5$ data points and consider a countable set $\mathcal{A}$ of size 2, such as $\{0, 1\}$ with the choice of $\alpha = 0.01$. We set the maximum length $L_{max} = 3$ that can be reliably used from formula 5.3.3. We examine the reliability of setting L_{max}, by checking that the hidden Markov Model inferred using CSSR with $L_{max} = 3$ is consistent with the model structure inferred using $L_{max} + 1 = 4$ [54]., i.e. the process has converged.

5.5 Results

Results for the depressing synapse are shown in Figure 5.3 and the histories of each causal state for all these machines are shown in Table 5.6.3 in Appendix 2. On the histograms, we indicated the maximum statistical complexity threshold with a red line. We can see that for low mean firing rates, the probability of getting large response values, "1", is high and its corresponding reconstructed CSM captures the dynamic with only one state. Similarly, for very high mean firing rate the probability of getting small response values, "0", is high and a one state machine results with the probabilities reversed. For intermediate mean firing rates, near the maximum entropy value of 2-3 Hz, the machine has 2 states, indicating more complex structure. Both 2 and 5 Hz result in identical machines in structure with slight variations in the transition probabilities. Note that even though the distribution sloshes around as the frequency is varied, there is little change in complexity in the epsilon machines through this range.

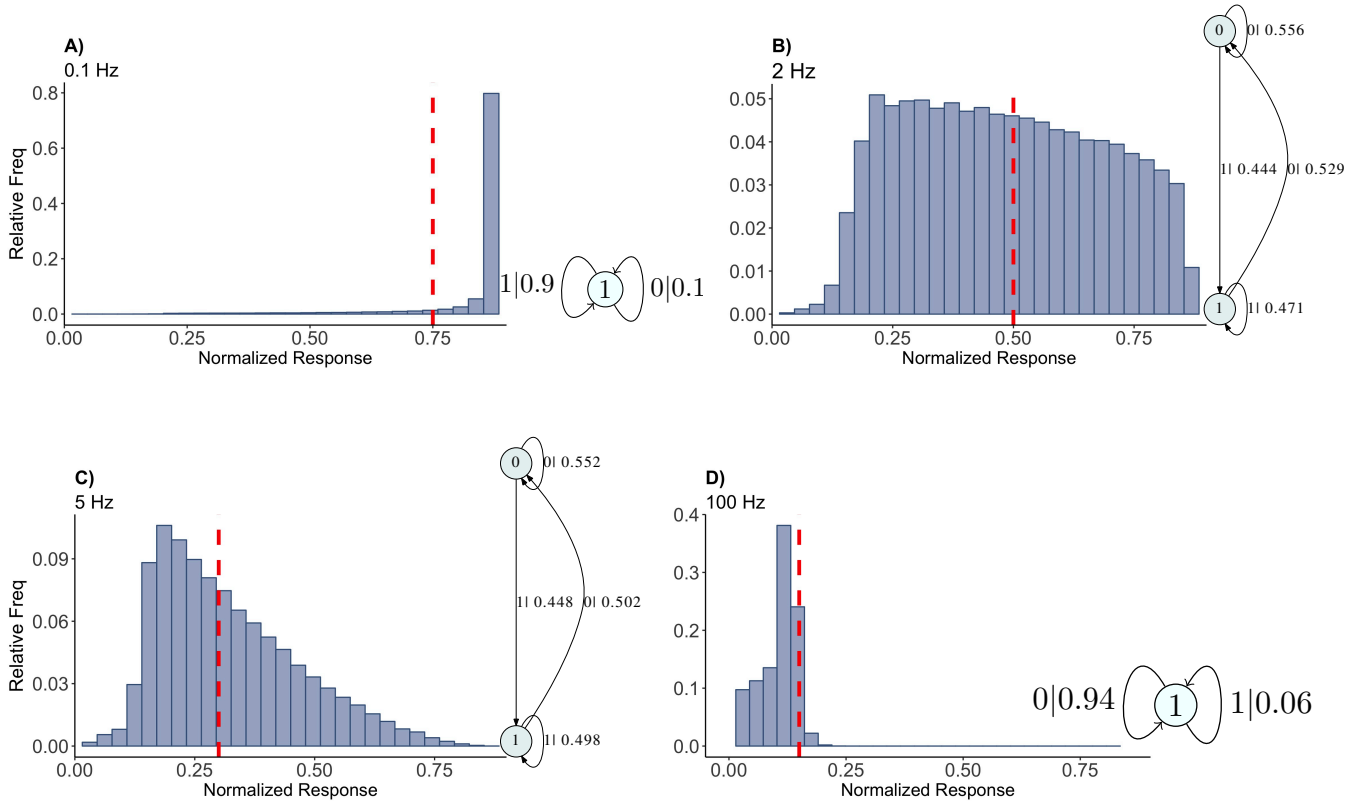


Figure 5.3: Causal state machines (CSMs) reconstructed and their corresponding relative frequency distributions obtained from depressing FD model. Model is stimulated by Poisson spike trains with mean firing rates **(A)** 0.1, **(B)** 2, **(C)** 5 and **(D)** 100 Hz. The transitions between states are indicated with symbol emitted during the transition (1= large synaptic response, 0 = small synaptic response) and the transition probability. In both **(A)** and **(D)**, CSMs for 0.1 and 100 Hz Poisson spiking process consist of a single state "1" which transitions back to itself, emitting a large response with probabilities 0.9 and 0.06 for low and very high mean firing rates, respectively. In both **(B)** and **(C)**, 2-state CSMs reconstructed for 2 and 5 Hz Poisson spiking process emit large responses with nearly similar probabilities.

Histograms of output response for the facilitating synapse are shown in Figure 5.4, for 50, 77, 100, 125, 200, 250 Hz along with their corresponding reconstructed CSMs of $L_{max} = 3$. all the machines can be described by referring to a persistent inner cycle and outer cycle. At

50 and 77 Hz, the machine structures are similar, with small variations in their transition probabilities. However, the histograms for these two mean firing rates are not similar and thus the machine identifies the underlying unifying stochastic process. At 50 and 77 Hz, the outer cycle connects state 0 to 1 to 2 and back to 0 while the inner cycle connects states 1 to 3 to 4 to 2 and back to 1. Note that an additional transition exists between state 3 and 2. State 4 has self-connecting edge that emits a "1". The self-connecting edge appears in all the other machines. At 100 Hz, the machine structure is more complex with more edges, vertices, and one set of parallel edges from state 3 to 6. This increase in complexity is not surprising as this is inflection point of the normalized fixed point response, See Figure 5.1. The machine found at 125 Hz has an outer cycle connecting states 0 to 1 to 3 and back to 0, while the inner cycle connects 1 to 2 to 3 and back to 1. The 200 Hz machine has the same number of inner cycles as 125 Hz machine. The machine for 250 Hz is the simplest and can be derived from the machine at 125 Hz by merging state 1 and 2.

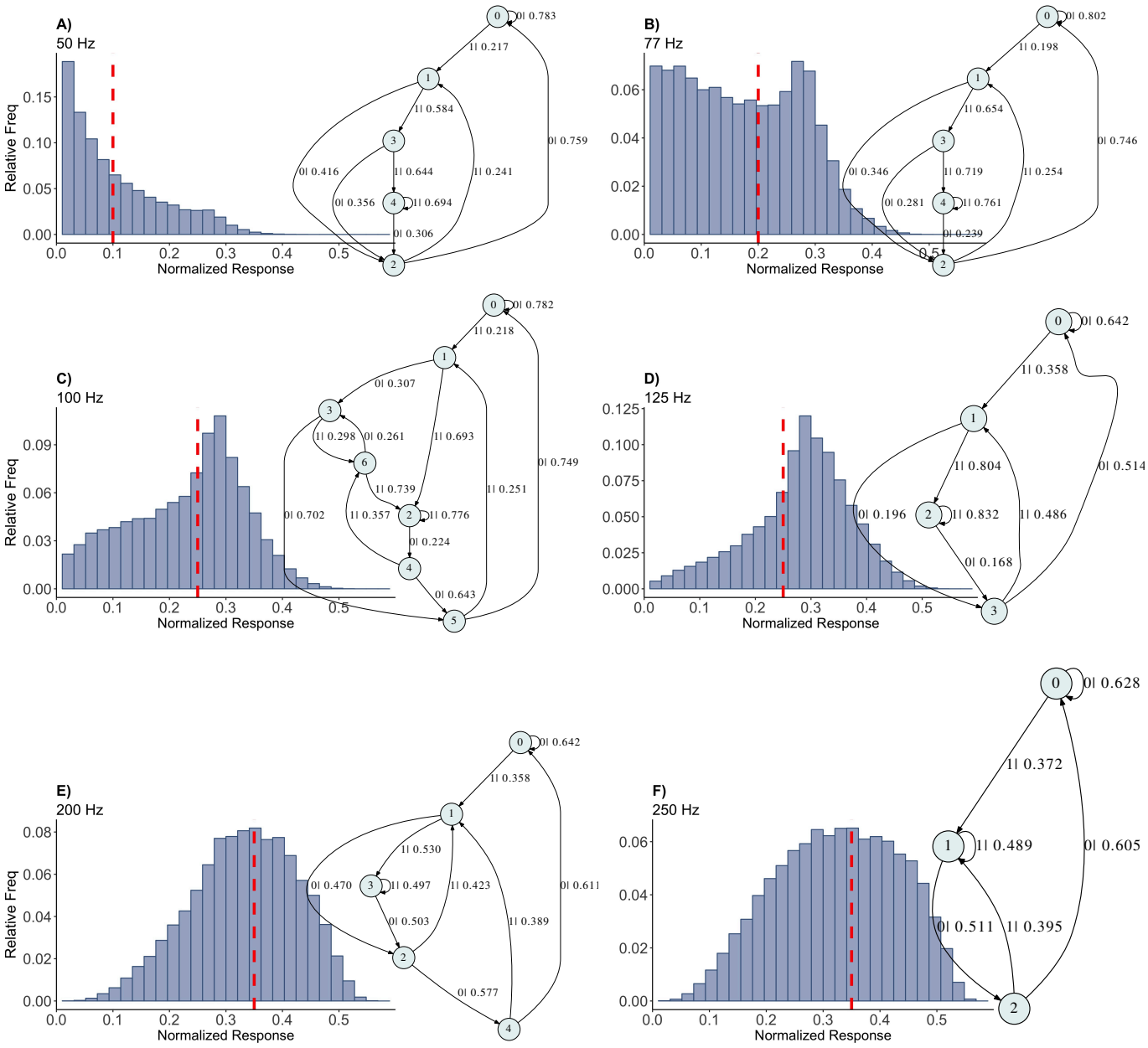


Figure 5.4: Causal state machines (CSMs) reconstructed and their corresponding relative frequency distributions obtained from facilitating FD model driven by Poisson spike train with mean firing rates **(A)** 50, **(B)** 77, **(C)**100, **(D)** 125, **(E)** 200 and **(F)** 250 Hz. State "0" is the baseline state. Similar graph structure is seen for mean firing rates of 50 and 70 Hz. Under mean firing rate of 100 Hz, the graph structure is more complex with more edges, vertices, and one set of parallel edges from state "3" to "6". In nonphysiological range from 125 to 250 Hz, the complexity of graph structure decreases.

The mixed synapse dynamic has a less complex structure compared to the facilitating synapse, see Figure 5.5. At 25 Hz, where local maximum occurs, the machine is complex. The machines at 5, 50, and 125 Hz have 2 states with small variations in the transition probabilities. At 25 and 250 Hz the machines are similar with different transition probabilities.

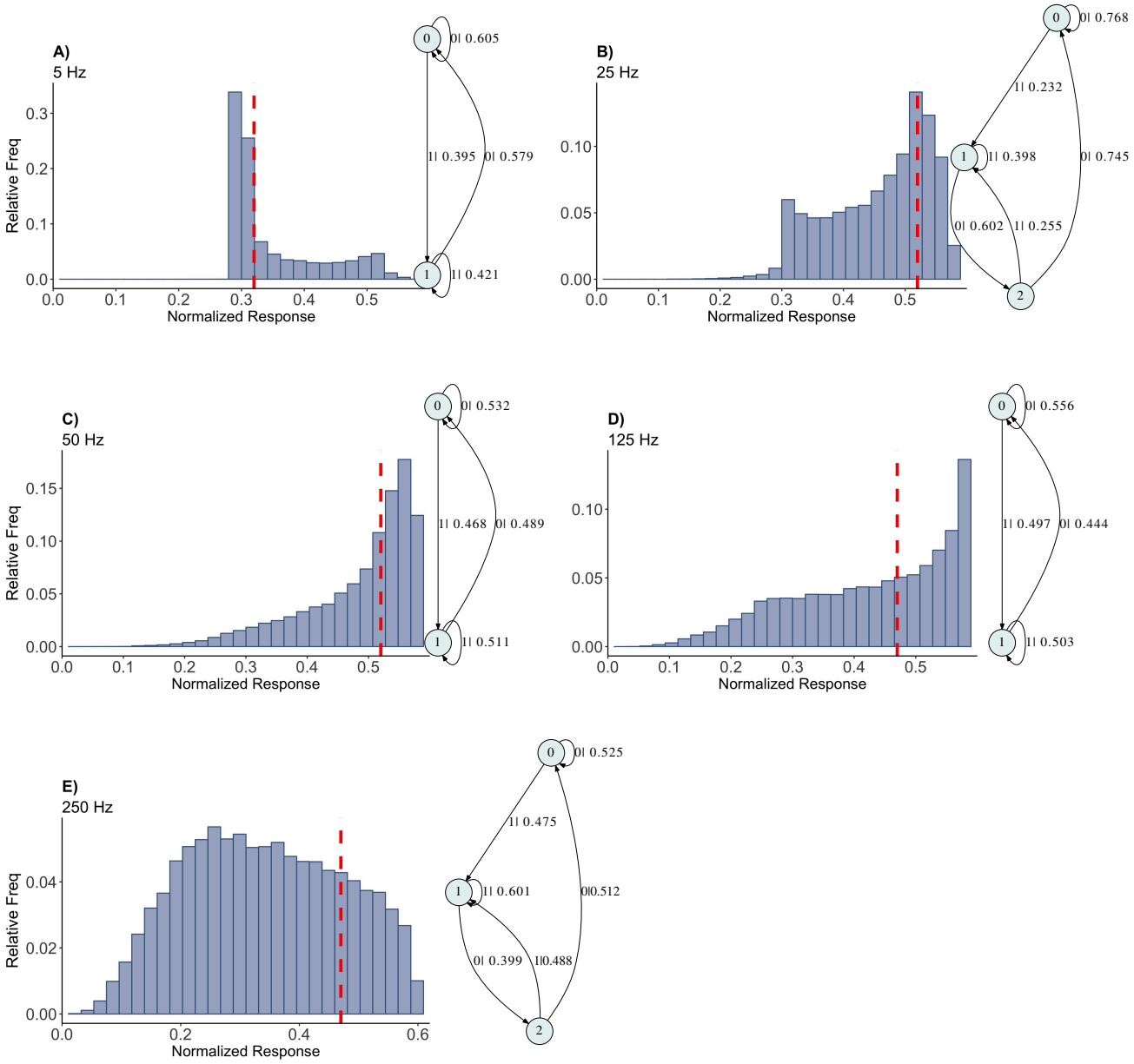


Figure 5.5: Causal state machines (CSMs) reconstructed and their corresponding relative frequency distributions obtained from mixed FD model driven by a Poisson spike train with mean rates **(A)** 5, **(B)** 25, **(C)** 50, **(D)** 125 and **(E)** 250 Hz. In **(A)**, **(C)**, and **(D)**, CSMs for mean firing rates of 5, 50, and 125 Hz consist of two states with similar structure, emitting successive large responses followed by small responses. 3-State CSM for mean firing rate 25 Hz has more complex graph structure. Note that this is inflection point for this synapse model, (see Figure 5.1).

5.6 Discussion

The depressing synapse has the simplest structure of the three cases. For very small and very large mean firing rates, the data points are mostly 1's and 0's, respectively, which results in one-state machines. For intermediate mean firing rates the transition from left to right skewness of the response distribution occurs, the 2-state CSM explains the dynamics. The structural content of the response in facilitating and mixed synapses are more complex. The complexity of the machines for these three synapses can be understood by comparing the decomposed fixed point spectrum. Figure 5.2 clearly shows the difference between three cases. For instance, in the depressing synapse, the response fixed point P_r is completely controlled by changes in the fraction of readily releasable vesicles R, while in the facilitating synapse both the fraction of release ready vesicles and probability of release of vesicles changes across the spectrum. The mixed synapse has a small variation in release probability, while fraction of release ready vesicles changes across the spectrum. Therefore, both R and P play role in increasing the complexity of the machines, although in an indirect manner.

The goal of this chapter was to present methods for exploring the structural content of output response while making minimal a priori assumptions as to the form of that structure. It is however clear that this work needs to be seen as one of the first investigations into the application of computational mechanics in differentiating synapse types and that the number of limitations need to be addressed before large neuronal problems can be tackled. There is still work to be done regarding the implementation of the algorithm and the tunning of the parameters. For instance, we would like to implement a more reliable techniques to choose maximum history length, such as minimizing Schwartz's Bayesian Information Criterion (BIC) over history length. BIC is known to be consistent for selecting the order of Markov chains. Also, we plan in future work to investigate different choices of threshold other than traditional defaults as the selection of this value can have dramatic effects on model accuracy.

Appendix 1

Table 5.6.1, 5.6.2 and 5.6.3 shows the histories of length $L = 2$ or $L = 3$ of the causal states for mixed, facilitating and depressing synapse for different firing rates.

Table 5.6.1: The casual states of ϵ-machine reconstructed from mixed synapse data

	State name	Histories $\overleftarrow{x}$	Morph Pr(1|$\overleftarrow{x}$)	Pr(State)
rate=5 hz	0	00, 10, 000, 010, 100, 110	0.39497	0.59461
	1	01, 11, 001,011, 101, 111	0.42066	0.40538
rate=25 hz	0	00, 000, 100	0.23205	0.54651
	1	01, 11, 011, 001, 111, 101	0.39858	0.28318
	2	10, 010, 110	0.25531	0.17030
rate=50 hz	0	00, 10, 010, 000, 110, 100	0.46753	0.51145
	1	01, 11, 011, 001, 111, 101	0.51053	0.48855
rate=125 hz	0	11, 01, 111, 101, 001, 011	0.50224	0.47144
	1	10, 00, 110, 100, 000, 010	0.44397	0.52855
rate=250 hz	0	00, 000, 100	0.474059	0.546511
	1	01, 11, 011, 001, 111, 101	0.601587	0.28318
	2	10, 010, 110	0.488314	0.17030

Table 5.6.2: The casual states of ϵ-machine reconstructed from facilitating synapse data

	State name	Histories $\overleftarrow{x}$	Morph $\Pr(1\|\overleftarrow{x})$	Pr(State)
rate=50 hz	0	00, 000, 100	0.21658	0.47924
	1	01, 001, 101	0.58359	0.13662
	2	10, 010, 110	0.24030	0.13662
	3	1, 011	0.64431	0.07973
	4	11, 111	0.69382	0.16778
rate=77 hz	0	00, 000, 100	0.19802	0.44910
	1	01, 001, 101	0.65377	0.11920
	2	10, 010, 110	0.25396	0.11920
	3	1, 011	0.71936	0.077931
	4	11, 111	0.76098	0.23455
rate=100 hz	0	000	0.21851	0.27980
	1	001	0.69264	0.08160
	2	11, 011, 111	0.77652	0.39156
	3	010	0.29741	0.03601
	4	110	0.35668	0.08750
	5	100	0.25076	0.08159
	6	101	0.73926	0.04192
rate=125 hz	0	00	0.35821	0.17470
	1	01	0.80351	0.12185
	2	11	0.83166	0.58160
	3	10	0.48645	0.12184
rate=200 hz	0	00, 000	0.35822	0.21312
	1	01, 001, 101	0.52965	0.21672
	2	10, 010, 110	0.42328	0.21672
	3	11, 011, 111	0.49759	0.22846
	4	100	0.38921	0.12498
rate=250 hz	0	0, 00	0.37234	0.35489
	1	1, 01, 11	0.48872	0.42686
	2	10	0.39453	0.21824

Table 5.6.3: The casual states of ϵ-machine reconstructed from depressing synapse data

	State name	Histories $\overleftarrow{x}$	Morph $\Pr(1\|\overleftarrow{x})$	Pr(State)
rate= 0.1 hz	0	11, 10, 01, 00, 111, 110, 101, 100, 011, 010, 001, 000	0.902427	1
rate= 2 hz	0	00, 10, 000, 010, 100, 110	0.443366	0.544131
	1	01, 11, 001, 011, 101, 111	0.470802	0.455869
rate= 5 hz	0	11, 01, 111, 101, 001, 011	0.498163	0.471689
	1	10, 00, 110, 100, 000, 010	0.448049	0.528311
rate= 100 hz	0	10, 11, 00, 01, 100, 101, 110, 111, 000, 001, 010, 011	0.0582917	1

Appendix 2

The choice of the threshold impacts the results, therefore in order to examine the effect of thresholding on the structure of machines, we shifted the default thresholds for each type of short-term plasticity slightly downwards ($\tau = -0.05$) and upwards ($\tau = +0.05$). We observe that in depressing and mixed synapses in physiological range, small changes in the partitioning threshold does not affect the models' structure and they are robust within this small interval with slight variations in the transition probabilities. However, in facilitating synapse due to higher complexity in the structure of machines, we observed more variation in the structure of the CSMs with varying the threshold. These results are summarized in Tables 5.6.4, 5.6.5 and 5.6.6 and Figures 5.6, 5.7 and 5.8.

Table 5.6.4: Depressing synapse parameter set. Topology of the machines (number of states) and statistical complexity (SC) as threshold is varied up and down by 0.05 from the maximum statistical complexity value for four different input frequencies.

Firing rate (Hz) / Threshold (τ)/ Number of States/(SC)	0.1	2	5	100
τ/number of states/(SC)	0.75/1/(0)	0.5/2(0.996)	0.3/2/(0.998)	0.15/1/(0)
$\tau + 0.05$/number of states	0.8/1/(0)	0.55/2/(0.960)	0.35/2/(0.945)	0.2/1/(0)
$\tau - 0.05$/number of states	0.7/1/(0)	0.45/2/(0.995)	0.25/2/(0.965)	0.1/1/(0)

Table 5.6.5: Mixed synapse parameter set. Topology of the machines (number of states) and statistical complexity as threshold is varied up and down by 0.05 from the maximum statistical complexity value for four different input frequencies.

Firing rate (Hz) / Threshold (τ)/ Number of States/(SC)	5	25	50	125
τ/number of states/(SC)	0.32/2/(0.97)	0.52/3/(1.491)	0.52/2/(0.998)	0.47/2/(0.971)
$\tau + 0.05$/number of states	0.37/2/(0.85)	0.57/3/(1.091)	0.57/2/(0.934)	0.52/2/(0.670)
$\tau - 0.05$/number of states	0.27/2/(0.88)	0.47/3/(1.202)	0.47/2/(0.950)	0.42/2/(0.866)

Table 5.6.6: Facilitating synapse parameter set. Topology of the machines (number of states) and statistical complexity as threshold is varied up and down by 0.05 from the maximum statistical complexity value for four different input frequencies.

Firing rate (Hz) / Threshold (τ)/ Number of States/(SC)	50	77	100	125
τ/number of states/(SC)	0.1/5/(1.470)	0.2/5/(2.02)	0.25/7/(2.106)	0.25/4/(2.49)
$\tau + 0.05$/number of states	0.15/5/(1.410)	0.25/5/(1.184)	0.3/5/(1.711)	0.3/4/(2.266)
$\tau - 0.05$/number of states	0.05/5/(1.452)	0.15/5/(2.001)	0.2/5/(2.056)	0.2/4/(2.451)

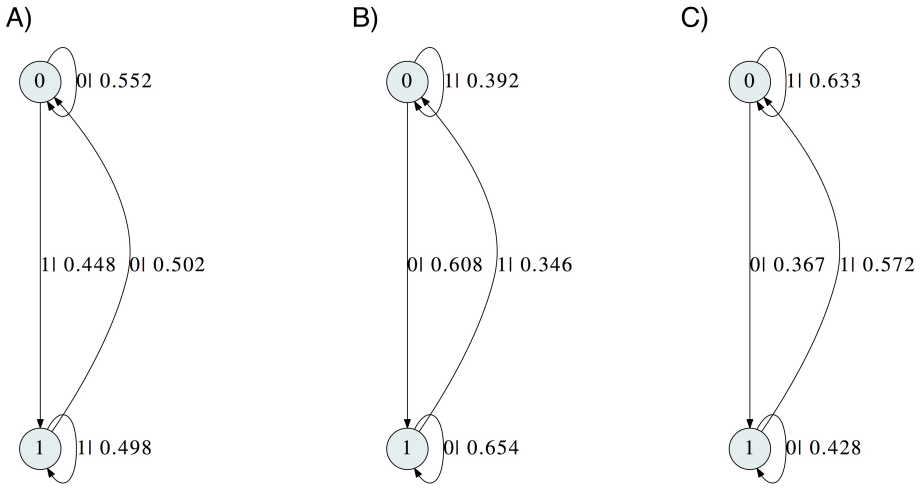


Figure 5.6: Depressing synapse parameter set. Causal state machines at input frequency of 5 Hz for varying partition threshold. The maximum statistical complexity threshold value, τ, is 0.3 (machine shown in A)), In B) and C) the machine for $\tau + 0.05 = 0.35$ and $\tau - 0.05 = 0.25$, respectively.

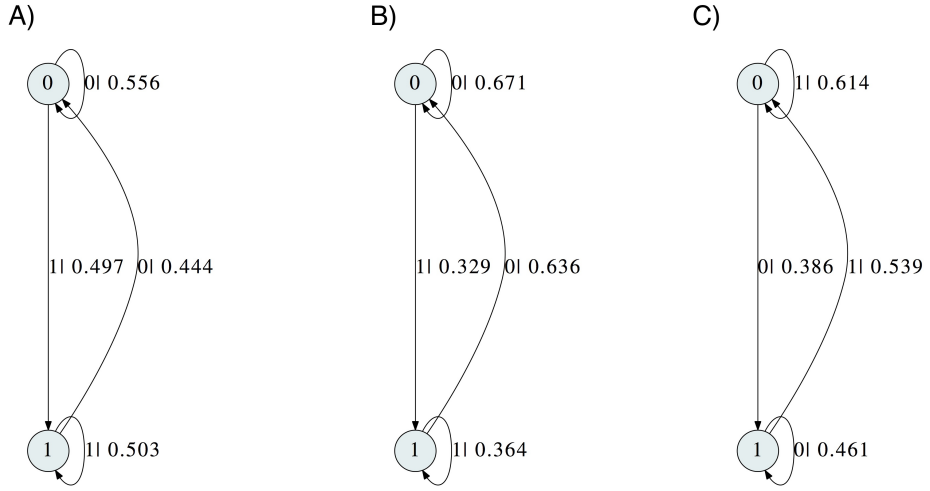


Figure 5.7: Mixed synapse parameter set. Causal state machines at input frequency of 25 Hz for varying partition threshold. The maximum statistical complexity threshold value, τ, is 0.52 (machine shown in A)), In B) and C) the machine for $\tau + 0.05 = 0.57$ and $\tau - 0.05 = 0.47$, respectively.

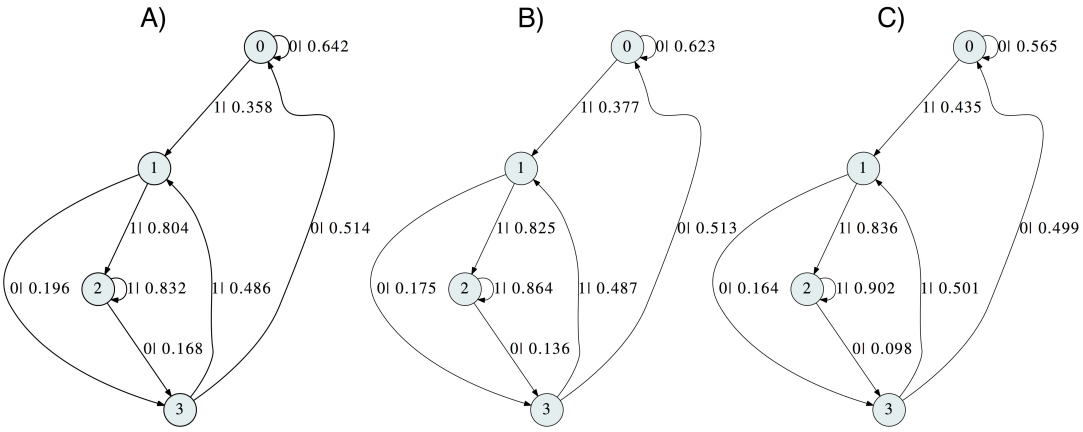


Figure 5.8: Facilitating synapse parameter set. Causal state machines at input frequency of 125 Hz for varying partition threshold. The maximum statistical complexity threshold value, τ, is 0.25 (machine shown in A)), In B) and C) the machine for $\tau + 0.05 = 0.3$ and $\tau - 0.05 = 0.2$, respectively.

Bibliography

[1] L. F. Abbott and Wade G. Regehr. Synaptic computation. *Nature*, 431:796–803, Oct 2004.

[2] E. Adrian. *Basis of Sensation*. New York. NY: Norton, 1928.

[3] E. D. Adrian and Rachel Matthews. The action of light on the eye. *The Journal of Physiology*, 63(4):378–414, 1927.

[4] E. D. Adrian and Yngve Zotterman. The impulses produced by sensory nerve-endings. *The Journal of Physiology*, 61(2):151–171, 1926.

[5] Andrs Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(34):163–193, 2001.

[6] Jennifer Boes and Charles R. Meyer. Multi-variate mutual information for registration. *Lect Notes Comput Sci*, 1679:606–612, 09 1999.

[7] M. Carl, S. Bangalore, and M. Schaeffer. *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. New Frontiers in Translation Studies. Springer International Publishing, 2015.

[8] BO CARTLING. Control of neural information transmission by synaptic dynamics. *Journal of Theoretical Biology*, 214(2):275 – 292, 2002.

[9] John D. Clements and R.Angus Silver. Unveiling synaptic plasticity: a new graphical and analytical approach. *Trends in Neurosciences*, 23(3):105 – 113, 2000.

[10] M.X. Cohen. *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press, 2014.

[11] Laura Lee Colgin. Rhythms of the hippocampal network. *Nature Reviews Neuroscience*, 17:239–249, Mar 2016. Review Article.

[12] Wikimedia Commons. File:blausen 0657 multipolarneuron.png — wikimedia commons, the free media repository, 2018. [Online; accessed 9-April-2018].

[13] Houghton Mifflin Company. Synapse. *The American Heritage ® Science Dictionary*, Apr 2018.

[14] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction To Algorithms*. MIT electrical engineering and computer science series. MIT Press, 2001.

[15] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[16] James P. Crutchfield. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1):11 – 54, 1994.

[17] James P. Crutchfield and Karl Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, Jul 1989.

[18] Peter Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2005.

[19] Jeremy S. Dittman and Wade G. Regehr. Calcium dependence and recovery kinetics of presynaptic depression at the climbing fiber to purkinje cell synapse. *Journal of Neuroscience*, 18(16):6147–6162, 1998.

[20] Daniel Dufresne. Algebraic properties of beta and gamma distributions, and applications. *Advances in Applied Mathematics*, 20(3):285 – 299, 1998.

[21] Paul Embrechts and Charles Goldie. Perpetuities and random equations. In Petr Mandl and Marie Hušková, editors, *Asymptotic Statistics*, pages 75–86, Heidelberg, 1994. Physica-Verlag HD.

[22] David Freedman and Persi Diaconis. On the histogram as a density estimator:l2 theory. *Probability Theory and Related Fields*, 57(4):453–476, December 1981.

[23] Galit Fuhrmann, Idan Segev, Henry Markram, and Misha Tsodyks. Coding of temporal information by activity-dependent synapses. *Journal of Neurophysiology*, 87(1):140–148, 2002.

[24] F. Gabbiani and S.J. Cox. *Mathematics for Neuroscientists*. Elsevier Science & Technology Books, 2017.

[25] Peter Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6):369 – 373, 1988.

[26] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. Dram: Efficient adaptive mcmc. *Statistics and Computing*, 16(4):339–354, Dec 2006.

[27] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 04 2001.

[28] James E. Hanson. *Computational Mechanics of Cellular Automata, PhD thesis*. University of California, Berkeley, 1993.

[29] G.T. Heineman. *Algorithms In A Nutshell*. Shroff Publishers & Distributors, 2008.

[30] Hanspeter Herzel and Ivo Grosse. Measuring correlations in symbol sequences. *Physica A: Statistical Mechanics and its Applications*, 216(4):518 – 542, 1995.

[31] Pawe Hitczenko. Convergence to type i distribution of the extremes of sequences defined by random difference equation. *Stochastic Processes and their Applications*, 121(10):2231 – 2242, 2011.

[32] Marc W. Howard and Howard Eichenbaum. Time and space in the hippocampus. *Brain Research*, 1621:345 – 354, 2015. Brain and Memory: Old Arguments and New Perspectives.

[33] David Kelly, Mark Dillingham, Andrew Hudson, and Karoline Wiesner. A new method for inferring hidden markov models from noisy time sequences. *PLOS ONE*, 7(1):1–9, 01 2012.

[34] Harry Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Math.*, 131:207–248, 1973.

[35] K. H. Knuth. Optimal data-based binning for histograms. *ArXiv Physics e-prints*, May 2006.

[36] L. F. Kozachenko and N. N. Leonenko. On statistical estimation of entropy of random vector. *Problems Infor. Transmiss.*, 23(2):95–101, 1987.

[37] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Probl. Inf. Transm.*, 23(1-2):95–101, 1987.

[38] A Kraskov. *Synchronization and Interdependence Measures and their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data*. PhD thesis, University of Wuppertal, NIC Series. : Research Center Juelich, 2004.

[39] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[40] J. Josh Lawrence, Heikki Haario, and Emily F. Stone. Presynaptic cholinergic neuromodulation alters the temporal dynamics of short-term depression at parvalbumin-positive basket cell synapses from juvenile ca1 mouse hippocampus. *Journal of Neurophysiology*, 113(7):2408–2419, 2015.

[41] Chuang-Chung J. Lee, Mihai Anton, Chi-Sang Poon, and Gregory J. McRae. A kinetic model unifying presynaptic short-term facilitation and depression. *Journal of Computational Neuroscience*, 26(3):459, Dec 2008.

[42] Joseph T. Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11, 2014.

[43] Wolfgang Maass and Henry Markram. Synapses as dynamic memory buffers. *Neural Networks*, 15(2):155 – 161, 2002.

[44] Donald M. MacKay and Warren S. McCulloch. The limiting information capacity of a neuronal link. *The bulletin of mathematical biophysics*, 14(2):127–135, Jun 1952.

[45] William J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, Jun 1954.

[46] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review*, 63(2):81–97, 1956.

[47] R. M. Mnatsakanov, N. Misra, Sh. Li, and E. J. Harner. Kn-nearest neighbor estimators of entropy. *Mathematical Methods of Statistics*, 17(3):261–277, Sep 2008.

[48] R. Moddemeijer. On estimation of entropy and mutual information of continuous distributions. *Signal processing*, 16(3):233–248, 1989.

[49] N.L. Novère. *Computational Systems Neurobiology*. Springer Netherlands, 2012.

[50] M. Palus. Identifying and quantifying chaos by using information-theoretic functionals. in a. s. weigend and n. a. gerschenfeld, editors. In *In A. S. Weigend and N. A. Gerschenfeld, editors, Time Series Prediction: Forecasting the Future and Understanding the Past, NATO Advanced Research Workshop on Compar- ative Time Series Analysis*, pages 387–413, Sante Fe, NM, May 1994. Addison-Wesley.

[51] Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, June 2003.

[52] Richard M. Restak. *The Brain: The Last Frontier*. Toronto: Bantam, 1984.

[53] Fred Rieke, Davd Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA, USA, 1999.

[54] J.M. Schwier, R.R. Brooks, C. Griffin, and S. Bukkapatnam. Zero knowledge hidden markov model inference. *Pattern Recognition Letters*, 30(14):1273 – 1280, 2009.

[55] DAVID W. SCOTT. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[56] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2009.

[57] Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3):817–879, Aug 2001.

[58] Cosma Rohilla Shalizi and Kristina Lisa Klinkner. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In Max Chickering and Joseph Y. Halpern, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)*, pages 504–511, Arlington, Virginia, 2004. AUAI Press.

[59] Cosma Rohilla Shalizi, Kristina Lisa Shalizi, and James P. Crutchfield. An algorithm for pattern discovery in time series. *CoRR*, cs.LG/0210025, 2002.

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

[61] C.E. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, 1963, 1949.

[62] Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23(3-4):301–321, 2003.

[63] Richard B. Stein. The information capacity of nerve cells using a frequency code. *Biophys J*, 7(6):797–826, Nov 1967. 19210999[pmid].

[64] Charles F. Stevens. Quantal release of neurotransmitter and long-term potentiation. *Cell*, 72(Supplement):55 – 63, 1993.

[65] Charles F Stevens and John F Wesseling. Activity-dependent modulation of the rate at which synaptic vesicles become available to undergo exocytosis. *Neuron*, 21(2):415 – 424, 1998.

[66] Emily Stone, Heikki Haario, and J Josh Lawrence. A kinetic model for the frequency dependence of cholinergic modulation at hippocampal gabaergic synapses. *Mathematical biosciences*, 258:162–75, 2014.

[67] Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

[68] G Tononi, O Sporns, and G M Edelman. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037, 1994.

[69] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company, 1977.

[70] Wim Vervaat. On a stochastic difference equation and a representation of non-negative infinitely divisible random variables. *Advances in Applied Probability*, 11(4):750–783, 1979.

[71] Jonathan D. Victor. Binless strategies for estimation of information from neural data. *Phys. Rev. E*, 66:051903, Nov 2002.

[72] Lu-Yang Wang and Leonard K. Kaczmarek. High-frequency firing helps replenish the readily releasable pool of synaptic vesicles. *Nature*, 394:384–388, Jul 1998.

[73] S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

[74] M. Wibral, R. Vicente, and J.T. Lizier. *Directed Information Measures in Neuroscience*. Understanding Complex Systems. Springer Berlin Heidelberg, 2014.

[75] Willfried Wienholt and Bernhard Sendhoff. How to determine the redundancy of noisy chaotic time series. *International Journal of Bifurcation and Chaos*, 06(01):101–117, 1996.

[76] Zhijun Yang, Matthias H. Hennig, Michael Postlethwaite, Ian D. Forsythe, and Bruce P. Graham. Wide-band information transmission at the calyx of held. *Neural Computation*, 21(4):991–1017, 2009. PMID: 19018705.

[77] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach*. Advances in Database Systems. Springer US, 2006.