Energy Efficiency of Neural Information Processing

A Dissertation

Presented to

the Faculty of the School of Engineering and Applied Science

University of Virginia

In Partial Fulfillment

of the requirements for the Degree

Doctor of Philosophy (Electrical Engineering)

by

Jie Xing

August 2014

© 2014 Jie Xing

APPROVAL SHEET

The dissertation is submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy (Electrical Engineering)

Jie Xing, Author

This dissertation has been read and approved by the examining committee:

Toby Berger, Dissertation Advisor

Stephen G. Wilson, Committee Chairman

Scott T. Acton

Christian Gromoll, Department of Mathematics

Terrence J. Sejnowski, Co-advisor, Salk Institute for Biological Studies

Accepted for the School of Engineering and Applied Science:

James H. Aylor, Dean, School of Engineering and Applied Science

August 2014

To my dear parents and beloved Sheen

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Toby Berger, for his excellent guidance, caring, understanding, patience and most importantly, persistent encouragement and unparalleled support during my graduate studies at the University of Virginia. His profound knowledge about information theory and probability theory, his enduring enthusiasm about neuroscience and his impeccable attitude towards research have greatly inspired me and helped me form a rigorous, dedicated, and creative research style for my future career life in the fields of information theory and neuroscience. I am also grateful to Dr. Terrence J. Sejnowski for his wonderful guidance and continuing support during my internship at the Salk Institute for Biological Studies and afterwards, which makes me stronger both physically and spiritually.

Besides my advisor, I would like to thank the rest of my PhD advisory committee: Professor Stephen G. Wilson, Professor Scott T. Acton, and Professor Christian Gromoll, for their support, encouragement, insightful comments, and valuable suggestions.

I am so blessed to be surrounded by many colleagues and friends at both the University of Virginia and the Salk Institute. I am particularly grateful to my friends and colleagues Ben S. Huang, Antonio Pinto Duarte, Maxime Bojean, Trygve Solstad, David Peterson, Mary Ellen Perry, Xin Wang for providing valuable suggestions to improve my Ph.D. studies. My sincere thanks also goes to my friends Zhenyu Guo, Jiaxing Guo, Tingjun Xie, Houbing Song and Chenguang Xu for their stimulating discussions and help, and for making my life at UVa enjoyable.

I also gratefully appreciate the financial support provided by NSF under grant CCF-1162449.

Finally, I would like to thank my parents, Juying Qiao and Yongliang Xing, for giving birth to me at the first place and unconditionally supporting me throughout my life. Last but not least, I would like to thank my girlfriend, Yu-sheen Wu, who was always there cheering me up and standing by me through good times and bad times.

Jie Xing

Charlottesville, Virginia

Abstract

The energy efficiency of information processing in the human brain is astonishingly superior to that of any machine yet designed by mankind. It is estimated that the 10^{11} neurons composing the human brain consume on average 20 watts of power, whereas the recent "real-time" simulation of some 10 million neurons in the cat visual cortex (headed by IBM Almaden Research Center) was 10^9 times more energy costly per neuron. Brains have evolved to prodigiously compute and communicate information with remarkable efficiency. Since neurons are expressly designed to exchange information with one another, it is fundamental to understand information processing and energy expenditure at the nodal level of the network. Furthermore, a steadily increasing fraction of neuroscientists subscribe to the view that each neuron's design should maximize the ratio between the rate at which it conveys information and the rate at which it expends energy. For all of the above stated reasons, my doctoral research explores the single-neuron modeling of information processing and energy efficiency from both theoretical and experimental perspectives.

The overall goal of this thesis is to analyze the performance of a single neuron, the smallest working unit of the brain, from an information-energy efficiency perspective. In particular, using information theory, random Poisson measures, Laplace transforms, and calculation of variations, we propose a mathematical framework for the stochastic processing and transmission of information performed at the neuronal level. We find the optimum distribution that characterizes the afferent excitatory/inhibitory postsynaptic potential (EPSP/IPSP) intensity by maximizing the Shannon mutual information rate given a constraint on the

total energy that a neuron expends for metabolism, postsynaptic potential generation, and action potential propagation during one interspike interval (ISI). This optimum distribution of the incoming EPSP/IPSP intensity serves as a bridge that specifies how an energy efficient brain needs to match the long-term statistics of each of its neuron's inputs to that neuron's particular design. Note that bits per Joule (bpJ) measures the performance of a neuron when viewed as a communication channel, since bits vs joule = bits/sec vs joule/sec = information rate vs power is the standard tradeoff considered by information theorists when studying a channel's capacity. We treat this tradeoff both analytically and through computational simulations for a series of increasingly sophisticated models.

In collaboration with the Salk Institute, we have tested the validity of this information-energy optimizing hypothesis using *in vivo* recordings of the visual thalamus from the cat. The experimentally-obtained statistical histograms are a close fit with the theoretically-derived optimum distributions. Imposing a bpJ-maximizing condition on single neuron function not only allows us to obtain key analytical conclusions that are in good agreement with experimental observations but also yields an intriguing bridge between single neuron theory and the theory of real neural networks, perhaps paving the way to wider applications in neuroscience and engineering. Overall, this research is a step further in the endeavor of reverse engineering of the brain.

Contents

Contents **vii**

List of Figures x

List of Abbreviations xii

1 Introduction **1**

1.1 Research Motivation and Literature Overview 1

1.2 Information Transmission between Neurons 2

1.3 Dissertation Outline 5

2 Theoretical and Computational Neuron Modeling: Unequal Synaptic Weight Model **7**

2.1 Overview 8

2.2 Problem Formulation and Preliminaries 10

2.3 Nature of the Randomness of Weight Vectors 11

2.4 Information Rate Calculation 13

2.5 Finding $f_{T_-|\Lambda}(t|\lambda)$: Mixtures of Gamma Distributions 17

2.5.1 Case A: Excitatory Synaptic Weights with Exponential Distribution . 18

2.5.2 Case B: Excitatory and Inhibitory Synaptic Weights with Gaussian Distribution 19

2.6 T_- is Gamma Distributed 20

2.7 From the Integral Equation to the Differential Equation 23

2.8 Summary 27

3 Theoretical and Computational Neuron Modeling: Gaussian Diffusion Model **29**

3.1 Problem Statement and Preliminaries 30

3.1.1 Neuron Outputs and Inputs 30

3.1.2 The Definition of Λ 31

3.2 Mutual Information and Energy Consumption 33

3.2.1 Information Rate . . . 34
3.2.2 Energy Expenditure . . . 35
3.3 The GIG and IG Probability Densities . . . 36
3.3.1 IG and GIG in the Neuroscience Literature . . . 37
3.3.2 The IG Neuron Channel Model . . . 37
3.3.3 Λ-Dependent Thresholds . . . 39
3.4 Mathematical Analysis of the IG Neuron Channel with $\mu = 1/2$ . . . 40
3.4.1 Finding $f_{T-\Delta}(t)$ when $\alpha = -1/2$ and $\mu = 1/2$ . . . 41
3.5 Evaluating the Needed Inverse Laplace Transform . . . 42
3.6 The Bits/Joule Maximizing $f_\Lambda(\lambda)$ . . . 43

4 Theoretical and Computational Neuron Modeling: Non-Gaussian Diffusion Model 45

4.1 Problem Statement and Preliminaries . . . 46
4.1.1 Introduction . . . 46
4.2 Neuronal Stimulation . . . 47
4.2.1 GIG Hitting Times . . . 49
4.2.2 Modeling Neural Inputs and Outputs . . . 52
4.2.3 Information Rate . . . 54
4.3 What is being optimized inside the neuron? . . . 55
4.4 Bits per Joule Optimality Condition . . . 56
4.5 Marginal Output ISI Distribution $f_T(t)$ is a GIG distribution . . . 58
4.6 Shannon Mutual Information $I(\Lambda;T)$ Calculation . . . 61
4.7 Energy Expenditure $e(\Lambda,T)$ Calculation . . . 66
4.8 Possible Explanations for $I(\Lambda;T) = e(\Lambda,T)$ . . . 67
4.9 Bits per Joule Optimizing Input Probability Distribution $f_\Lambda(\lambda)$ . . . 69
4.10 Summary . . . 71

5 Experimental and Biophysical Neuron Modeling: The Paradoxical Energy Efficiency of Retinothalamic Transmission 72

5.1 Problem Statement . . . 72
5.2 Berger-Levy Energy Efficient Theory of Neural Computation and Communication 75
5.3 Statistical Analysis . . . 77
5.4 Methods Summary . . . 80
5.4.1 Theoretical Analysis . . . 81
5.4.2 LGN Neuron Model . . . 83

5.4.3 Two-to-four Fold Reasoning . . . 87
5.4.4 Mutual Information Estimation of the Spike Trains . . . 90
5.4.5 Optimal and Suboptimal Spike Trains Demonstration . . . 91
5.4.6 Power Law Distributed RGC Arrival Numbers in Each LGN ISI . . . 93
5.4.7 Comparison between Two Point Estimators of Parameter m . . . 94
5.4.8 Detailed Figure Captions . . . 95

6 Conclusions and Future Work **98**
6.1 Conclusions . . . 98
6.2 Future Work . . . 100
6.2.1 KKT instead of Lagrange Multipliers . . . 100
6.2.2 What Do Neurons Do and Not Do Highly Energy Efficiently? . . . 101
6.2.3 Network Coding . . . 102

Bibliography **103**

List of Figures

1.1 Illustration of neuronal signaling. [28] 3

1.2 Illustration of single synapse. [29] 4

2.1 Single interspike interval (ISI) schematic with the illustration of all physical parameters. 9

4.1 GIG densities for four choices of parameters (a, b, c) in Eq. (4.10). 50

4.2 Outline of an elephant with the wiggling trunk. [51] 52

4.3 GIG Channel 1. 62

4.4 GIG Channel 2. 62

4.5 Information vs energy plot. 67

4.6 $f_\Lambda(\lambda)$ plots for two sets of $(\alpha, \beta, \gamma, a, b, c)$. 70

5.1 Retinothalamic transmission. 73

5.2 LGN neuron model. 76

5.3 Thalamic interspike interval (ISI) parameter estimation and retinal averaged firing rate parameter estimation. 78

5.4 Best fit parameters of model to data for optimizing information/energy. . . 80

5.5 Retinal averaged firing rate parameter estimation from four typical data sets. 84

5.6 Detailed thresholding algorithm and computational simulation results. . . . 86

5.7 Thalamic spike timing estimated from four typical data sets. The histograms of spike timing (blue) are compared with the predicted ones from the LGN neuron model (red). 87

5.8 Parametric spaces of bits per joule optimized average number of arrivals per ISI (m_{max}), mutual information estimation (I), information per energy (I_j), and information per second (I_s). 89

5.9 *In vivo* and *in silico* optimal and suboptimal spike trains demonstration. . . 91

5.10 Histogram of the number of RGC arrivals in each LGN ISI from one typical data set. 92

5.11 Demonstration of arriving EPSPs in consecutive ISIs. 93

6.1 Illustration of neurons learning mathematical statistics and information theory. 99

List of Abbreviations

AP	action potential
ATP	adenosine triphosphate
B-L or BL	Berger-Levy
bpJ	bits per joule
bps	bits per second
cdf	cumulative distribution function
CI	confidence interval
EPSP	excitatory postsynaptic potential
GIG	generalized inverse Gaussian
IG	inverse Gaussian
IIF	idealized integrate-and-fire
i.i.d.	independent and identically distributed
IPSP	inhibitory postsynaptic potential
ISI	interspike interval
KKT	Karush-Kuhn-Tucker
LGN	lateral geniculate nucleus
LIF	leaky integrate-and-fire
LTD	long-term depression

LTP	long-term potentiation
ML	maximum likelihood
MVUE	minimum variance unbiased estimator
pdf	probability density function
PSP	postsynaptic potential
RGC	retinal ganglion cell
r.v.	random variable
SD	standard deviation
SNR	signal-to-noise ratio
V1	primary visual cortex

Chapter 1

Introduction

Reality is merely an illusion, albeit a very persistent one.

-Albert Einstein

1.1 Research Motivation and Literature Overview

Neuronal information processing is energetically costly. The energy efficiency of information processing in the human brain is astonishingly superior to that of any machine yet designed by mankind to address similar problems. It is estimated that the 10^{11} neurons composing the human brain consume on average 20 watts of power [1] [2], whereas the recent "real-time" simulation of some 10 million neurons in the cat visual cortex headed by IBM Almaden Research Center was 10^9 times more costly per neuron [3]. Energy supply restrictions on information processing have driven brains to evolve towards computing and communicating information with remarkable efficiency. Indeed, energy minimization subject to functional constraints is widely believed to be a potential unifying principle in understanding neuronal function [4]. Since neurons are evolutionarily designed to exchange information with one another, it is fundamental to understand information processing and energy expenditure at the nodal level of neuronal networks. Furthermore, it is essential to understand how each neuron regulates its information-processing efficiency by maintaining an adequate supply

of adenosine triphosphate (ATP), the molecular currency for energy transfer. Moreover, a steadily increasing fraction of neuroscientists are subscribing to the view that a neuron's optimal design should maximize the ratio between the rate of its information processing and the rate of its energy expenditure. For all of the above reasons, my doctoral research aims to explore single neuron modeling of information and energy processing from both theoretical and experimental perspectives. To better comprehend neuronal information processing and communication from an information-energy standpoint, we proposed to study mathematical models of single neurons as engines of computation and communication based on the homogeneous Poisson process [5] [6] [7], the Wiener process [8] and a more general Lévy diffusion process [9] [10] [11].

Information theory has often been applied to neuroscientific data analysis and biological systems modeling [12] [13] [14] [15]. However, energy-efficient neural codes have been studied for less than thirty years [16] [17] [18]. Evidence supporting energy efficiency principles has been reported for ion channels [4], action potentials [19], synapses [20], photoreceptors [21], in the retina [22] [23], grey matter [24], white matter [25] and in cortex [26]. In their highly regarded tutorial paper, Laughlin and Sejnowski discussed communication in cortical networks from an energy-efficiency point of view and emphasized the importance of energy efficiency [27].

1.2 Information Transmission between Neurons

A neuron in the brain receives information in the form of individual spike trains from its neighboring neurons and/or from peripheral sensory ganglia called the neuron's *afferent cohort*. Each of these spikes, also known as action potentials (AP's), arrives at a unique connection between the axonal terminals of an afferent (presynaptic) neuron and the dendritic tree or the soma of the neuron we seek to analyze. These AP's propagate through presynaptic axons and trigger the release of neurotransmitters that cross the synapses to elicit a postsynaptic

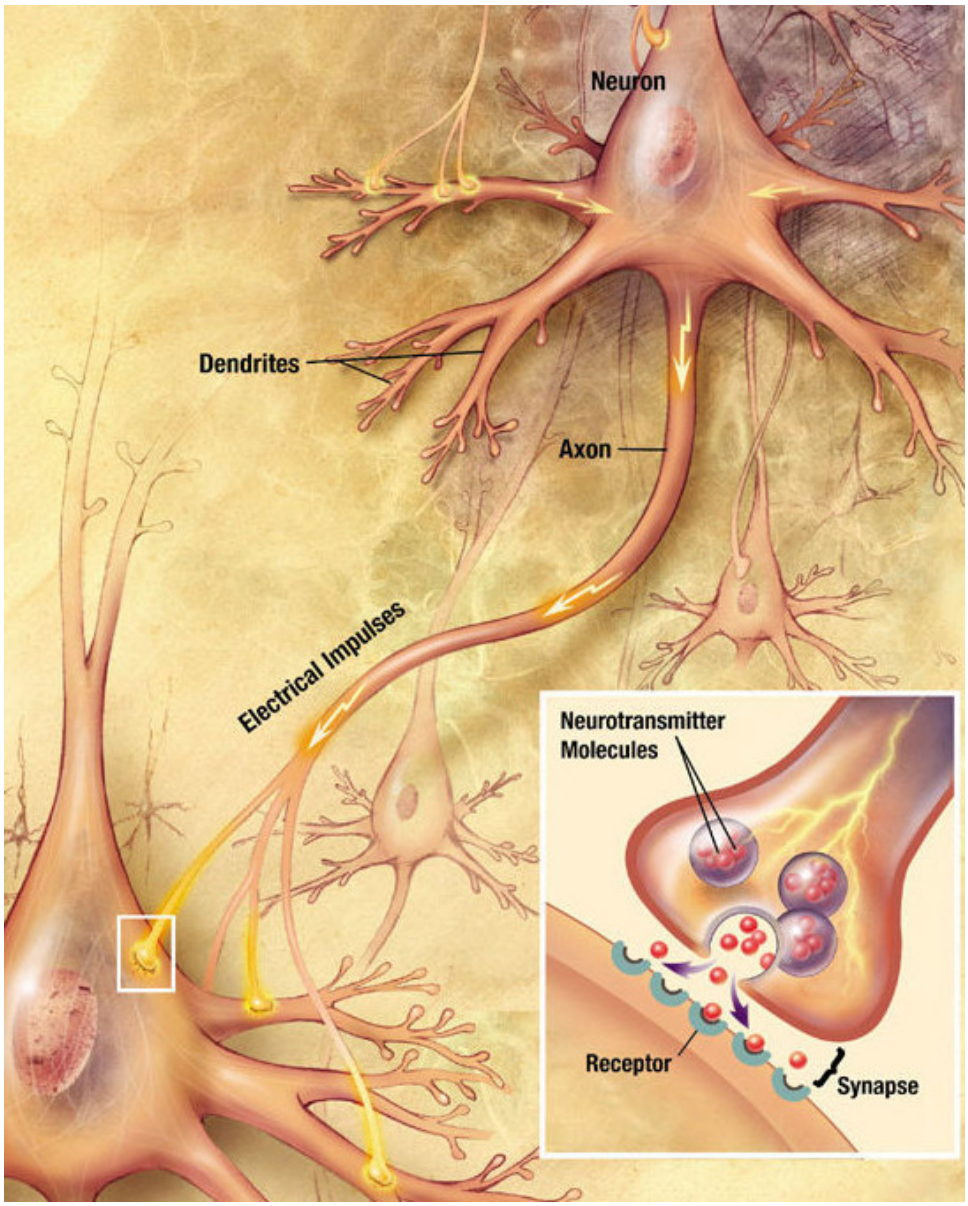

Figure 1.1: Illustration of neuronal signaling. [28]

potential in the postsynaptic cell. Once the accumulation of excitatory postsynaptic potential (EPSPs) exceeds a given threshold of membrane voltage, a new action potential is triggered in the postsynaptic neuron and the relay of information continues. The durations of the interspike intervals (ISI's) between a neuron's successive AP's constitute a random sequence which can be denoted by $\{T_k, k = 0, 1, 2, \dots\}$. We henceforth call that neuron "neuron j", or just "j" when that is unambiguous.

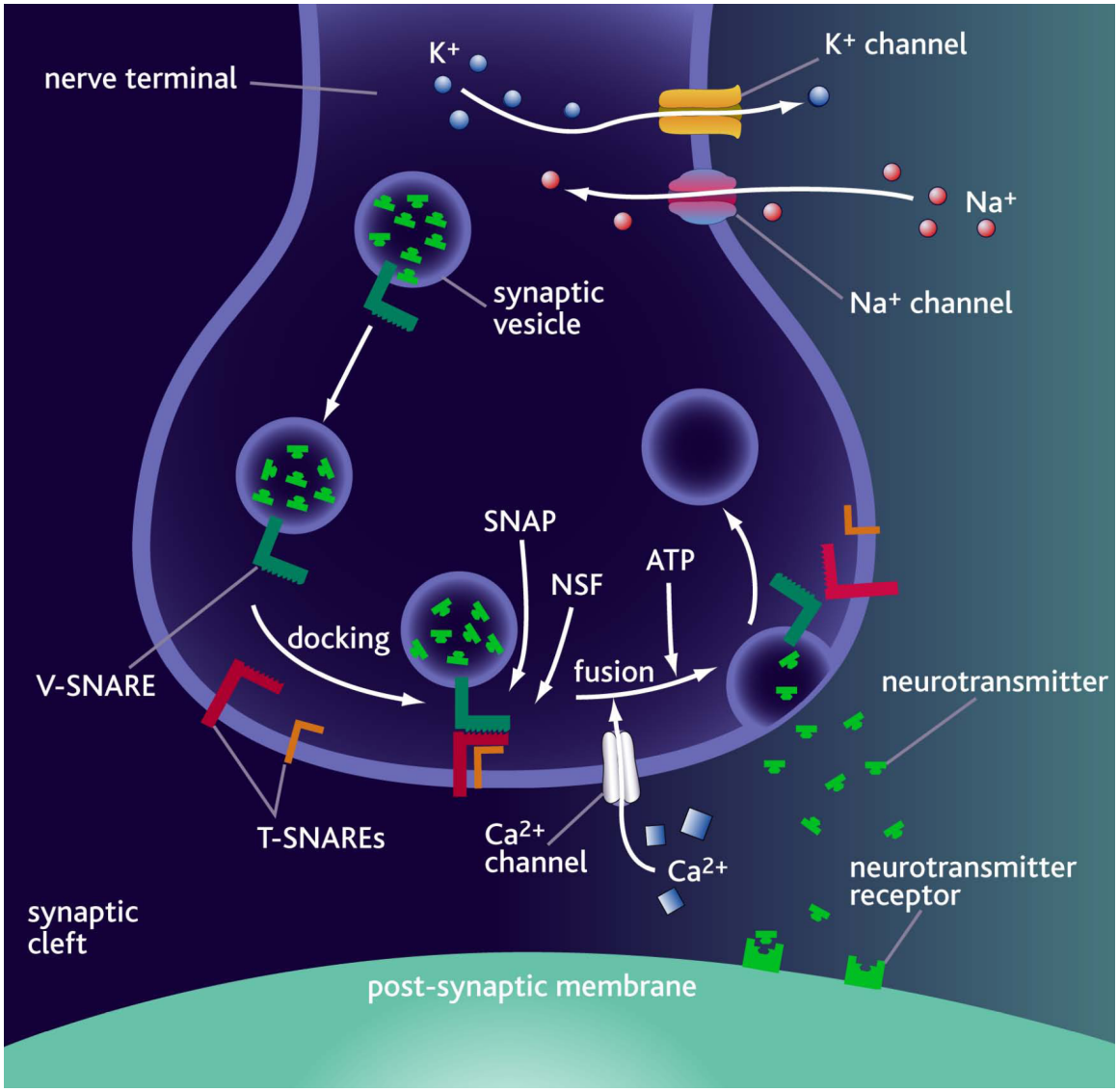

Figure 1.2: Illustration of single synapse. [29]

Neuron j's external excitation comprises a collection of neural spike trains generated by the members of j's afferent cohort. When a presynaptic spike arrives at one of j's presynaptic terminals, it elicits a transient membrane depolarization that opens voltage-gated calcium channels in the presynaptic bouton and drives a rapid influx of calcium. In response to the increase in calcium, vesicles containing neurotransmitters then fuse with the presynaptic membrane and release the neurotransmitters into the synapse. Those neurotransmitters diffuse across the synaptic cleft and bind to the postsynaptic receptors, which in turn trigger

the opening of ion channels and subsequent ionic influx resulting in a change of voltage across the postsynaptic membrane, e.g. an excitatory postsynaptic potential (EPSP). Each time the change in membrane potential exceeds a time-varying threshold value, j generates another output spike that propagates along its axon to the set of neurons referred to as j's efferent cohort or its targets. A typical neuron in the primate sensory cortex possesses circa $10,000$ inputs and circa $10,000$ outputs.

Each time j emits an efferent spike, there ensues a refractory period during which j cannot produce another spike. Neuroscientists distinguish two types of refractory periods - absolute and relative. Limiting attention to absolute refractoriness permits us to assume with negligible error that j's refractory periods all have the same duration, denoted by Δ. During each refractory period j strives to replenish the supply of chemicals in its various compartments in anticipation of producing a subsequent spike in response to the ongoing afferent excitation. However, j's state (i.e., the chemical concentrations in all of its subcompartments) does not always return to the same value at the end of each refractory period. In particular, in the not infrequent case of j experiencing several successive ISI's whose durations are only slightly larger than Δ, diminished supplies may prevent full restoration of the concentration of certain chemical species in some of j's subcompartments before the refractory period ends.

1.3 Dissertation Outline

The remainder of this dissertation is organized as follows:

In Chapter 2, a mathematical framework is formulated, based on homogeneous Poisson processes, for describing how a single neuron stochastically processes and communicates information. The neuron possesses both excitatory and inhibitory synapses with differing weights. Moreover, input and output distributions of the modeled neuron and long-term mutual information rate are derived.

In Chapter 3, a mathematical framework, based on inverse Gaussian processes, for how a single neuron stochastically processes and conveys information is studied. Again, input and output distributions of the modeled neuron are presented.

In Chapter 4, a mathematical framework, based on generalized inverse Gaussian processes, for how a single neuron stochastically computes and communicates information is presented. Once more, mutual information rate and energy consumption are calculated and analyzed.

In Chapter 5, we apply our energy efficient neuroscientific framework to analyze experimental *in vivo* recordings of the visual thalamus of the cat and resolve the paradoxical energy efficiency of retinothalamic transmission.

In Chapter 6, we recapitulate the major conclusions of this dissertation and provide some directions for future research.

The major results have been documented and published in conference proceedings and peer-reviewed journal papers.

Chapter 2

Theoretical and Computational Neuron Modeling: Unequal Synaptic Weight Model

Everything should be made as simple as possible, but not simpler.

-Albert Einstein

How neurons in the cerebral cortex process and transmit information is a long-standing question in systems neuroscience. To understand the neural mechanism from an information-energy efficiency standpoint, Berger and Levy calculated the maximum Shannon mutual information transfer per unit of energy expenditure of an idealized integrate-and-fire (IIF) neuron whose excitatory synapses all possess the same weight. In this chapter their IIF model is extended to a biophysically more realistic one in which synaptic weights not only are no longer assumed equal but also can be inhibitory. Using information theory, random Poisson measures, Laplace transforms and the maximum entropy principle, it is established that the probability density function (pdf) of interspike interval (ISI) duration induced by the bits per joule (bpJ) maximizing pdf $f_\Lambda(\lambda)$ of the combined excitatory and inhibitory postsynaptic potential (EPSP/IPSP) intensity remains a delayed gamma distribution as

in the IIF model. It is also shown that, in the case of unequal weights, $f_\Lambda(\cdot)$ satisfies an inhomogeneous Cauchy-Euler equation with variable coefficients for which the general solution form is provided.

2.1 Overview

The human brain, only two percent of the body's weight, accounts for twenty percent of the body's energy consumption [1] [2]. It is estimated that the circa 10^{11} neurons composing the human brain expend on average 20 watts of power, whereas the "real-time" simulation of some 10 million neurons in cat visual cortex headed by IBM Almaden Research Center in 2009 expended 10^9 times more energy per neuron than the cat does [3]. Brains have evolved that prodigiously compute and communicate information with remarkable efficiency. Energy minimization subject to functional constraints may be a unifying principle [4] [5]. Since neurons are expressly designed to exchange information with one another, understanding how information processing and energy expenditure are performed at the nodal level of the network is fundamental for acquiring insights into the limits of energy efficiency in computation and communication.

Information theory has been widely employed in neuroscientific data interpretation and system modeling during the last fifty years [12] [13] [14] [15]. However, energy efficient neural codes have been studied for less than two decades [16] [17] [18]. Evidence for energy efficiency has been reported for ion channels [4], photoreceptors [21], retina [22] [23], and cortex [24] [26]. Laughlin and Sejnowski discussed the cortex as a communicating network from an energy efficiency perspective [27]; Mitchison et al. and Chklovskii et al. applied energy efficiency to analyze cortical wiring and brain mapping [30] [31]; Berger and Levy proposed an energy efficient mathematical model for information transmission by a single neuron [5].

We seek to analyze the performance of a single neuron, the smallest working unit of the brain, from an information-energy efficiency perspective. By maximizing the Shannon mutual

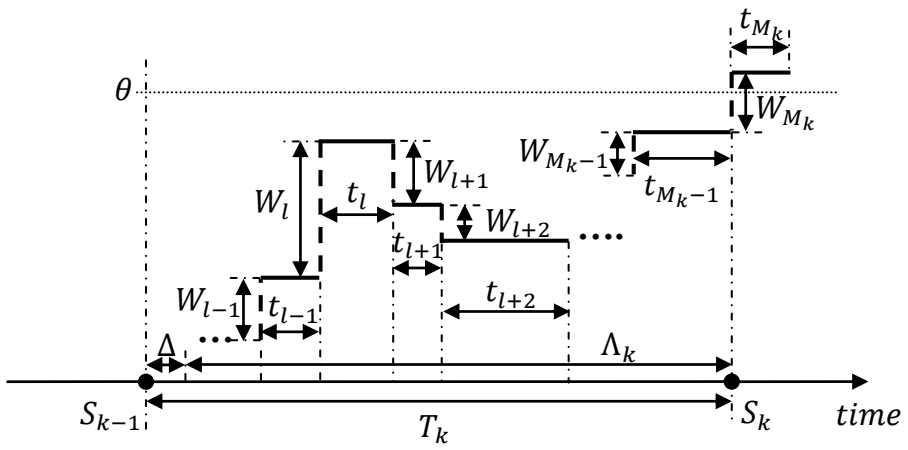

Figure 2.1: Single interspike interval (ISI) schematic with the illustration of all physical parameters.

information rate subject to a constraint on the total energy cost that a neuron expends for metabolism, postsynaptic potential generation, and action potential propagation during one interspike interval (ISI), Berger and Levy found that the reciprocal of afferent excitatory postsynaptic potential (EPSP) intensity, $1/\Lambda$, and interspike interval, T, follow a beta and a gamma distributions, respectively [5]. The primary goal of this Chapter is to extend this energy efficient model such that synapses can have unequal weights and can be inhibitory.

Our main and most striking result is that ISI durations continue to be gamma distributed even when synaptic weights need not be equal and may have either algebraic sign. Also, we show that the energy-constrained capacity-achieving combined excitatory/inhibitory postsynaptic potential (EPSP/IPSP) intensity for the case of unequal synaptic weights is no longer beta distributed as its fixed-weight counterpart, but instead is the solution of an inhomogeneous Cauchy-Euler equation with variable coefficients.

2.2 Problem Formulation and Preliminaries

We first introduce a mathematical framework for how a single neuron stochastically processes and communicates information. We consider a neuron in the primary sensory cortex that we refer to as "neuron j" or simply j for short. Let $\underline{W}_k = (W_1, W_2, \ldots, W_{M_k})$, where W_l is the weight of lth excitatory/inhibitory synapse of j to receive a spike during the kth ISI and produce an EPSP/IPSP in response thereto. We time-order the EPSP's & IPSP's according to the times at which they arrive at the somatic membrane (more specifically, at j's axon initial segment that connects its soma to its axon) and hence contribute to j's postsynaptic potential (PSP) accumulation. M_k is the integer-valued random cardinality of $\underline{W}_k$.

We assume that the components of $\underline{W}_k$ are chosen independent and identically distributed (i.i.d.) according to a certain cumulative distribution function (cdf) $F_W(w) = P[W \leq w], w \in (-\infty, +\infty)$. We model the lth contribution to j's PSP accumulation to be $W_l \cdot u(t - t_l)$; here, W_l is a random variable (r.v.) with the aforementioned cdf $F_W(w)$ and $u(t - t_l)$ equals 1 for $t \geq t_l$ and equals 0 for $t < t_l$. [1] We continue to assume as in [5] that:

- The PSP spiking threshold is constant at the value θ.
- Synaptic events that occur during any of the refractory periods that immediately follow j's spiking instants make no contribution to j's PSP.
- Each of j's refractory periods has the same duration, Δ.

Although we permit negative weights, we nonetheless require $\mathbb{E}[W] > 0$ in order to ensure that the threshold eventually is reached.

We model the EPSP/IPSP's production times in response to spikes from j's afferent cohort as an inhomogeneous Poisson measure with instantaneous rate function, $A(t)$, defined by

$$A(t) := \lim_{\mathrm{d}t \to 0} \frac{P[\text{one EPSP/IPSP arrival in}(t, t + \mathrm{d}t)]}{\mathrm{d}t}. \tag{2.1}$$

[1] Real neurons do not exhibit ideal step function EPSP/IPSP responses to afferent synaptic events. However, their responses are sharply rising/decaying functions that usually complete more than 90% of their rise/decay during less than 0.1 ms, so an idealized step function response is not far off the mark.

Then as in [5] we take a time average operation over the rate function $A(t)$ and obtain

$$\Lambda_k := \frac{1}{T_k - \Delta} \int_{S_{k-1}+\Delta}^{S_k} A(u)\mathrm{d}u, \tag{2.2}$$

where Δ is the duration of j's refractory period, T_k is the kth ISI duration of j and $S_k = T_1 + T_2 + \cdots + T_k$.

Henceforth, we suppress the ISI index k and just write $\underline{W}$, M, T and Λ, and assume that when $\Lambda = \lambda$, EPSP/IPSP's are generated during the ISI according to a homogeneous Poisson point process with intensity λ.

Here we are interested in the Shannon mutual information, $I(\Lambda; T)$. Although this has been defined for a single pair of r.v.'s Λ and T, it is shown in Section 2.4 that it is a good first-order approximation to the long-term average mutual information rate in bits per spike, namely

$$I := \lim_{n\to\infty} \frac{1}{n} I(\Lambda_1, \Lambda_2, \ldots, \Lambda_n; T_1, T_2, \ldots, T_n), \tag{2.3}$$

lacking only an information decrement that addresses correlation among successive Λ_i's [5] [7].

Define $T_- := T - \Delta$. Since T_- is a one-to-one function of T, we have $I(\Lambda; T) = I(\Lambda; T_-)$, which in turn is defined [32] [33] [34] as

$$I(\Lambda; T_-) = \mathbb{E} \log \left[\frac{f_{T_-|\Lambda}(T_-|\Lambda)}{f_{T_-}(T_-)} \right], \tag{2.4}$$

where the expectation is taken with respect to the joint distribution of Λ and T_-.

Toward determining $I(\Lambda; T_-)$, we proceed to analyze $f_{T_-|\Lambda}(t|\lambda)$ and $f_{T_-}(t)$ with unequal synaptic weights.

2.3 Nature of the Randomness of Weight Vectors

Even if the excitatory synaptic weights of neuron j were known, $\underline{W} = (W_1, W_2, \ldots, W_M)$ would still be random because the time-ordered vector $\underline{R}$ of synapses excited during an ISI is

random. However, for purposes of mathematical analysis of neuron behavior it is not fruitful to restrict attention to a particular neuron with a known particular set of synaptic weights. Rather, it is more useful to think in terms of the histogram of the synaptic weight distributions of neurons in whatever neural region is being investigated. When many such histograms have been ascertained, if their shapes almost all resemble one another closely, then they can be arithmetically averaged to obtain a population histogram with fine resolution in the weights of its synaptic bins. This, in turn, would permit one to approximate this fine histogram by a continuous amplitude probability distribution of synaptic weights. (Then the analysis becomes more widely applicable than if one were to have used the exact weights of a particular neuron, especially considering that even that neuron will have a different set of weights in the future because of ongoing synaptic modification.) Moreover, the strong similarity of synaptic weight distributions has been observed through experiments [35]. Therefore, in the analysis that follows we take the view that the components of $\underline{W}$ are selected randomly from this continuous amplitude probability distribution. Said random distribution of synaptic weights also incorporates the random number of neurotransmitter-containing vesicles that are released when a spike is afferent to the synapse, the random number of excitatory/inhibitory neurotransmitter molecules in these vesicles, how many of those cross the synaptic cleft, bind to receptors and thereby generate EPSP's/ISPS's.

This model of random selection of weights comprising $\underline{W}$ is applicable both to ISI's in which the afferent firing rate Λ is large and to those in which it is small. When the value λ assumed by Λ is large, $\underline{W}$'s components just get selected more rapidly than when λ is small, but they continue to come from the same distribution. This implies that the expected number of them in a single ISI remains the same. Hence, from now on, we assume that the weight vector components $(W_1, W_2, \ldots, W_M)$ are jointly independent of Λ.

Neurons vary their synaptic weights in order to learn and adapt to changing environments. The synaptic plasticity of weights is widely considered essential to learning and memory [35]. Long term applicability of our model can be achieved by adjusting the weight distribution

$F_W(\cdot)$. When synaptic plasticity changes the overall weight distribution substantially, then $F_W(\cdot)$ should be changed accordingly. In this manner our theoretical model can connect neural plasticity to the communication of information by the neuron, thereby broadening its applicability.

2.4 Information Rate Calculation

We assume that information is only subadditive over disjoint ISIs [5], i.e., the inputs Λ_k's are not i.i.d.; in particular when Λ_k is considerably greater than $\mathbb{E}[\Lambda]$, then it is highly likely that the same is true for Λ_{k+1}. Therefore, the long-term mutual information rate per ISI, I, at which j's efferent spiking times $\{T_k\}$ provide information about j's random excitation over an ongoing sequence of ISI's, upon our having adopted the "mean value assumption" in Eq. (2.2), is defined as below:

$$I = \lim_{n\to\infty} \frac{1}{n} I(\underline{\Lambda}; \underline{T}) \tag{2.5}$$

$$= \lim_{n\to\infty} \frac{1}{n} I(\Lambda_1, \Lambda_2, \ldots, \Lambda_n; T_1, T_2, \ldots, T_n) \tag{2.6}$$

$$= \lim_{n\to\infty} \frac{1}{n} [I(\Lambda_1, \Lambda_2, \ldots, \Lambda_n; T_1, T_2, \ldots, T_{n-1}) + I(\Lambda_1, \Lambda_2, \ldots, \Lambda_n; T_n | T_1, T_2, \ldots, T_{n-1})]. \tag{2.7}$$

Next, we can reduce the first term in Eq. (2.7), sans the limit and average over n, as follows:

$$\begin{aligned} &I(\Lambda_1, \Lambda_2, \ldots, \Lambda_n; T_1, T_2, \ldots, T_{n-1}) \\ =&I(\Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1}; T_1, T_2, \ldots, T_{n-1}) + I(\Lambda_n; T_1, T_2, \ldots, T_{n-1} | \Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1}). \end{aligned} \tag{2.8}$$

Due to the fact that given $(\Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1})$, current average spiking intensity Λ_n is independent of historical ISI sequence $(T_1, T_2, \ldots, T_{n-1})$, which implies the second term in

Eq. (2.8) equals zero.

Meanwhile, the second term inside the limit in Eq. (2.7) can be written as [5]:

$$\begin{aligned} & I(\Lambda_1, \Lambda_2, \ldots, \Lambda_n; T_n | T_1, T_2, \ldots, T_{n-1}) \\ =& I(\Lambda_n; T_n | T_1, T_2, \ldots, T_{n-1}) + I(\Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1}; T_n | \Lambda_n, T_1, T_2, \ldots, T_{n-1}). \end{aligned} \tag{2.9}$$

Assuming as in [5] that the communication channel from $\{\Lambda_k\}$ to $\{T_k\}$ is memoryless with causal feedback, when Λ_n is given, T_n is independent of both $(\Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1})$ and $(T_1, T_2, \ldots, T_{n-1})$, which implies that the second term in Eq. (2.9) is zero and accordingly the first term in Eq. (2.9) can be derived as following:

$$I(\Lambda_n; T_n | T_1, T_2, \ldots, T_{n-1}) = h(T_n | T_1, T_2, \ldots, T_{n-1}) - h(T_n | \Lambda_n, T_1, T_2, \ldots, T_{n-1}) \tag{2.10}$$

$$= h(T_n | T_1, T_2, \ldots, T_{n-1}) - h(T_n | \Lambda_n) \tag{2.11}$$

$$= I(\Lambda_n; T_n) - I(T_n; T_1, T_2, \ldots, T_{n-1}) \tag{2.12}$$

Combining Eqs. (2.7), (2.8), (2.9) and (2.12), we have

$$\begin{aligned} I &= \lim_{n\to\infty} \frac{1}{n} I(\Lambda_1, \Lambda_2, \ldots, \Lambda_n; T_1, T_2, \ldots, T_n) \\ &= \lim_{n\to\infty} \frac{1}{n} [I(\Lambda_n; T_n) - I(T_n; T_1, T_2, \ldots, T_{n-1}) + I(\Lambda_1, \Lambda_2, \ldots, \Lambda_{n-1}; T_1, T_2, \ldots, T_{n-1})]. \end{aligned} \tag{2.13}$$

By mathematical induction, it is easy to rewrite Eq. (2.13) as

$$I = \lim_{n\to\infty} \frac{1}{n}\left[\sum_{k=1}^{n} I(\Lambda_k; T_k) - \sum_{k=1}^{n-1} I(T_{k+1}; T_1, T_2, \ldots, T_k)\right]. \tag{2.14}$$

Based on the definition of relative entropy [33], i.e., Kullback-Leibler distance, we have

$$\begin{aligned} I &= \lim_{n\to\infty} \frac{1}{n}\left[\sum_{k=1}^{n} I(\Lambda_k; T_k) - \sum_{k=1}^{n-1} I(T_{k+1}; T_1, T_2, \ldots, T_k)\right] \\ &= \lim_{n\to\infty} \frac{1}{n}\left[\sum_{k=1}^{n} I(\Lambda_k; T_k) - D(f_{T_1,T_2,\ldots,T_n} || f_{T_1} \cdot f_{T_2} \cdots f_{T_n})\right] \qquad (2.15) \\ &= \lim_{n\to\infty} \frac{1}{n}\left[\sum_{k=1}^{n} I(\Lambda_k; T_k) - \left(\sum_{k=1}^{n} h(T_k) - h(T_1, T_2, \ldots, T_n)\right)\right]. \qquad (2.16) \end{aligned}$$

Since both the random process $\{T_k\}$ and the joint process $\{(\Lambda_k, T_k)\}$ are strictly stationary, we may replace $I(\Lambda_k; T_k)$ with $I(\Lambda_1; T_1)$, make $h(T_k) = h(T_1)$ and then suppress the index to yield that

$$\begin{aligned} I =& I(\Lambda; T) - \left(h(T_1) - \lim_{n\to\infty} \frac{1}{n} h(T_1, T_2, \ldots, T_n)\right) \qquad (2.17) \\ :=& I(\Lambda; T) - I_{decr}. \qquad (2.18) \end{aligned}$$

Note that the third term in Eq. (2.17) is the differential entropy rate of the strictly stationary continuous amplitude r.v.'s T_k's extended from Theorem 3.5.1 in [34].

The results of the differential entropy rate, mathematically speaking, involve the assumption that all the finite dimensional joint distribution of the $\{T_k\}$ process has to be jointly absolutely continuous even though, physically speaking, this assumption is trivial to be satisfied. It is our strong opinion that the channel always has a joint density $f(T_k|\Lambda_k)$ with no atoms conditioned on any subset of Λ_k's, that is, for $\Lambda \in (0, \infty)$ and $T \in (0, \infty)$ it is sufficient to say that the $\{T_k\}$ process is absolutely continuous.

$$h(\{T\}) = \lim_{n\to\infty} \frac{1}{n} h(T_1, T_2, \ldots, T_n) \tag{2.19}$$

$$= \lim_{n\to\infty} \frac{1}{n} \left[\sum_{k=1}^{n} h(T_k | T_{k+1}, \ldots, T_n) \right] \tag{2.20}$$

$$\stackrel{a}{=} \lim_{n\to\infty} \frac{1}{n} \left[\sum_{k=1}^{n} h(T_1 | T_2, \ldots, T_{n-k+1}) \right] \tag{2.21}$$

$$\stackrel{b}{=} \lim_{n\to\infty} \frac{1}{n} \left[\sum_{k=1}^{n} h(T_1 | T_2, \ldots, T_k) \right], \tag{2.22}$$

where equality (a) and equality (b) are based on the following statements, due to the strict stationarity of $\{T_k\}$, that $\forall\ k \leq n < \infty$,

$$I(T_{n-1}; T_n) = I(T_1; T_2) \tag{2.23}$$

$$h(T_k | T_{k+1}, \ldots, T_n) = h(T_1 | T_2, \ldots, T_{n-k+1}). \tag{2.24}$$

Therefore, the information decrement term, I_{decr}, can be written as

$$I_{decr} = h(T_1) - h(\{T\}) \tag{2.25}$$

$$= I(T_1; T_2) + \Big(h(T_1 | T_2) - h(\{T\}) \Big) \tag{2.26}$$

We can see $I(T_1; T_2)$ as the first order approximation of the second term in Eq. (2.15), the second term in Eq. (2.26) being the dependency/memory measurement.

In order to achieve the highest energy efficiency, neuron j likely strives to reduce the degree of dependence among the ISI durations it transmits. However, it is not possible to eliminate the dependence between T_1 and T_2. This is because j's inputs Λ_1 and Λ_2 are dependent in a manner that j is unable to control. Since the channel model from Λ's to T's is memoryless, T_1 and T_2 are dependent as well, so $I(T_1; T_2) > 0$. The only way that j could generate independent r.v.'s to transmit to its targets would be to encode $T_1, T_2, \cdots, T_n$ into a vector,

the components of which are independent. However this is an inefficient tradeoff because it both increases the computational and storage requirements in the neuron and introduces undesirable latency. Equation (2.26) shows that if the T process were first-order Markov, then $I(T_1;T_2)$ would be the entire information decrement. Even this less stringent first order Markov requirement is computationally demanding and injects latency. It is probably the case that neurons are structured so as to produce $\{T_k\}$'s that are well approximated by low order Markov processes so that the second term in Eq. (2.26) has a much smaller magnitude than that of the first one.

Note that the first order approximation, $I(T_1;T_2)$, in Eq. (2.26) can be approximated by $\max\{0, -\kappa E \log T_-\} + C$ as in Appendix C of [5], where $T_- = T - \Delta$.

2.5 Finding $f_{T_-|\Lambda}(t|\lambda)$: Mixtures of Gamma Distributions

The contribution to j's PSP accumulation attributable to the lth afferent spike during an ISI will be assumed to be a randomly weighted step function $W_l \cdot u(t - t_l)$, where t_l is the time at which it arrives at the postsynaptic membrane. [2]

It follows that the probability $P_m := P(M = m)$ that exactly m PSP's are afferent to j during an ISI is

$$P_m = P(\max_m\{W_1 + W_2 + \cdots + W_{m-1}\} < \theta,\ W_1 + W_2 + \cdots + W_m \geq \theta). \tag{2.27}$$

[2] In practice, $u(t - t_l)$ needs to be replaced by a $g(t - t_l)$, where $g(\cdot)$ looks like $u(\cdot)$ for the first 15 or so ms but then begins to decay. This has no effect when λ is large because the threshold is reached before this decay ensues. For small-to-medium λ's, it does have an effect but that could be neutralized by allowing the threshold to fall with time in an appropriate fashion. There are several ways to effectively decay the threshold, one being to decrease the membrane conductance.

Next, we write

$$\begin{aligned} & P(t \le T_- \le t+dt | \Lambda = \lambda) \\ = & \sum_{m=1}^{\infty} P(t \le T_- \le t+dt, M = m | \Lambda = \lambda) \\ = & \sum_{m=1}^{\infty} P_m \cdot P(t \le T_- \le t+dt | \Lambda = \lambda, M = m), \end{aligned} \tag{2.28}$$

where Eq. (2.28) holds because of the assumption that the weight vector $\underline{W} = (W_1, W_2, \ldots, W_M)$ is independent of Λ, which implies its random cardinality, M, is independent of Λ.

It follows as in [5] [6] [7] that, given $M = m$ and $\Lambda = \lambda$, T_- is the sum of m i.i.d. exponential r.v.'s with parameter λ, i.e., a gamma pdf with parameters m and λ. Summing over all the possibilities of M and letting dt become infinitesimally small, we obtain

$$f_{T_-|\Lambda}(t|\lambda) = \sum_{m=1}^{\infty} P_m \cdot \frac{\lambda^m t^{m-1} e^{-\lambda t}}{(m-1)!} u(t). \tag{2.29}$$

It is impossible to determine P_m in the general case. However, we have been able to compute it exactly in Cases A and B discussed below.

2.5.1 Case A: Excitatory Synaptic Weights with Exponential Distribution

Suppose the components of the weight vector are i.i.d. and have the exponential pdf $\alpha e^{-\alpha w_i}, \forall\ w_i \ge 0$ with $\alpha > 0$. Then we know that $Y_m := W_1 + W_2 + \cdots + W_m$ has the gamma pdf

$$f_{Y_m}(y_m) = \frac{\alpha^m y^{m-1} e^{-\alpha y}}{(m-1)!} u(y).$$

Since there are only positive weights, Eq. (2.27) can be reduced to

$$\begin{aligned} P_m =& P(Y_{m-1} < \theta,\ W_m \geq \theta - Y_{m-1}) \\ =& \int_0^{\theta} f_{Y_{m-1}}(u)du \int_{\theta-u}^{\infty} f_{W_m}(v)dv \\ =& \frac{(\alpha\theta)^{m-1}}{(m-1)!} e^{-\alpha\theta}. \end{aligned} \tag{2.30}$$

Therefore, it follows from Eq. (2.29) that

$$\begin{aligned} f_{T_-|\Lambda}(t|\lambda) =& \sum_{m=1}^{\infty} \frac{(\alpha\theta)^{m-1}}{(m-1)!} e^{-\alpha\theta} \cdot \frac{\lambda^m t^{m-1} e^{-\lambda t}}{(m-1)!} u(t) \\ =& \lambda e^{-(\alpha\theta+\lambda t)} \sum_{k=0}^{\infty} \frac{(\alpha\theta\lambda t)^k}{(k!)^2} u(t). \end{aligned} \tag{2.31}$$

The summation in Eq. (2.31) equals $I_0(2\sqrt{\alpha\theta\lambda t})$ where I_0 is the modified Bessel function of the first kind with order 0 [36]. Note that, if the threshold is made to vary inversely with time, i.e., if $\theta = C/\lambda t$, where C is a constant, then the density assumes the form

$$f_{T_-|\Lambda}(t|\lambda) = \lambda C_0 e^{-\left(\frac{C_1}{\lambda t} + \lambda t\right)} u(t), \tag{2.32}$$

where C_0 is $I_0(2\sqrt{\alpha C})$ and C_1 is a positive constant αC.

2.5.2 Case B: Excitatory and Inhibitory Synaptic Weights with Gaussian Distribution

When taking inhibitory synaptic weights into account, by defining $Y_m := \sum_{i=1}^{m} W_i$ and $f_W(w) := \frac{\mathrm{d}F_W(w)}{\mathrm{d}w}$ and based on Eq. (2.27), we may write

$$\mathrm{P}_m = \Pr(\max_m\{Y_{m-1}\} < \theta,\ Y_m \geq \theta) \tag{2.33}$$

$$= \Pr\left(W_1 < \theta, \sum_{i=1}^{2} W_i < \theta, \cdots, \sum_{i=1}^{m-1} W_i < \theta, \sum_{i=1}^{m} W_i \geq \theta\right)$$

$$= \int_{-\infty}^{\theta} \mathrm{d}w_1 f_W(w_1) \cdot \int_{-\infty}^{\theta - w_1} \mathrm{d}w_2 f_W(w_2) \cdots \int_{-\infty}^{\theta - \sum_{i=1}^{m-2} w_i} \mathrm{d}w_{m-1} f_W(w_{m-1}) \cdot \int_{\theta - \sum_{i=1}^{m-1} w_i}^{\infty} \mathrm{d}w_m f_W(w_m). \tag{2.34}$$

Suppose the components of the weight vector are i.i.d. and normally distributed, i.e.,

$$f_W(w_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(w_i-\mu)^2}{2\sigma^2}}, w_i \in (-\infty, +\infty), \tag{2.35}$$

where $\mu > 0$ and σ are the mean and the standard deviation of the Gaussian weight r.v. W.

Therefore, armed with Eqs. (2.34) and (2.35), we can always recursively calculate $\mathrm{P}_m, \forall m = 1, 2, 3, \cdots$ and then, using Eq. (2.29), to $f_{T_-|\Lambda}(t|\lambda)$.

2.6 T_- is Gamma Distributed

For the conditional pdf $f_{T_-|\Lambda}(t|\lambda)$ as in Eq. (2.29), letting $X = \lambda \cdot T_-$, we have the following equality in which $x = \lambda t$:

$$|f_{X|\Lambda}(x|\lambda)dx| = |f_{T_-|\Lambda}(t|\lambda)dt|.$$

It follows, in view of Eq. (2.29), that

$$f_{X|\Lambda}(x|\lambda) = \sum_{m=1}^{\infty} P_m \cdot \frac{x^{m-1}e^{-x}}{(m-1)!},\ \forall\ x \geq 0. \tag{2.36}$$

Note that λ not only doesn't explicitly appear on the right-hand side of Eq. (2.36) but also does not appear there implicitly within any of the P_m's; this is because, as noted earlier, M is independent of Λ, so P_m cannot be λ-dependent. Accordingly,

$$f_{X|\Lambda}(x|\lambda) = f_X(x) = \sum_{m=1}^{\infty} P_m \cdot \frac{x^{m-1}e^{-x}}{(m-1)!}, \ \forall\ x \geq 0. \tag{2.37}$$

Hence, although X equals $\Lambda \cdot T_-$, X nonetheless is independent of Λ. [3] We can rewrite the relationship as

$$T_- = \frac{1}{\Lambda} \cdot X, \tag{2.38}$$

where X is marginally distributed according to Eq. (2.37).

Then by taking logarithms in Eq. (2.38), we have

$$\log T_- = -\log \Lambda + \log X, \tag{2.39}$$

We see that Eq. (2.39) describes a channel with additive noise that is independent of the channel input. Specifically, the output is $\log T_-$, the input is $-\log \Lambda$, and the additive noise is $N := \log X$, which is independent of Λ (and therefore independent of $-\log \Lambda$) because X and Λ are independent of one another.

The mutual information between Λ and T_- thus is

$$\begin{aligned} I(\Lambda; T_-) &= I(\log \Lambda; \log T_-) \\ &= h(\log T_-) - h(\log T_- | \log \Lambda) \\ &= h(\log T_-) - h(N). \end{aligned} \tag{2.40}$$

[3] A simple example in which $X = AB$ is independent of A may be enlightening here. Let $P(A = -1) = P(A = 1) = 1/2$, $P(B = -2) = P(B = -1) = P(B = 1) = P(B = 2) = 1/4$, and assume A and B are independent. Note that, given either $A = -1$ or $A = 1$, X is distributed uniformly over $\{-2, -1, 1, 2\}$, so X is independent of A. X is not independent of B in this example because $|B| = 2$ implies $|X| = 2$, whereas $|B| = 1$ implies $|X| = 1$. In our neuron model the two factors of X, namely Λ and $T - \Delta$, are not independent of one another but rather are strongly negatively correlated.

Letting $Z = \log T_-$, we have

$$h(\log T_-) = h(Z) = -\mathbb{E} \log f_Z(Z).$$

Since $f_Z(z) = f_{T_-}(t) \cdot |dt/dz| = f_{T_-}(t) \cdot t$, it finally follows that

$$\begin{aligned} h(\log T_-) &= -\mathbb{E}[\log(f_{T_-}(T_-) \cdot T_-)] \\ \mathbb{E} &= h(T_-) - \mathbb{E} \log T_-. \end{aligned} \tag{2.41}$$

Thus, after substituting Eq. (2.41) into Eq. (2.40) and adding the information decrement term as in [5] [6] [7], the long-term average mutual information rate, I, we seek to maximize over the choice of $f_\Lambda(\cdot)$ is

$$I = h(T_-) + (\kappa - 1)\mathbb{E} \log T_- - h(N) - C. \tag{2.42}$$

From Equations (14-16) in [5], we see that the total expected energy expended is $e = C_1 \mathbb{E}T + C_2 \mathbb{E}M + C_3$. Since $\mathbb{E}T = \mathbb{E}T_- + \Delta$, $\mathbb{E}T$ can be converted to $\mathbb{E}T_-$ by adding $C_1\Delta$ to C_3, that is $C_3' := C_3 + C_1\Delta$. Hence, the optimization problem becomes

$$\begin{aligned} \underset{f_\Lambda(\lambda)}{\text{maximize}} \quad & I = h(T_-) + (\kappa - 1)\mathbb{E} \log T_- - h(N) - C \\ \text{subject to} \quad & e = C_1 \mathbb{E}T_- + C_2 \mathbb{E}M + C_3'. \end{aligned} \tag{2.43}$$

Since N and M are independent of Λ, the choice of $f_\Lambda(\cdot)$ affects I and e in Eq. (2.43) only through $f_{T_-}(\cdot)$ via the following equation

$$\int_0^\infty d\lambda f_\Lambda(\lambda) f_{T_-|\Lambda}(t|\lambda) = f_{T_-}(t), \ \forall \ 0 \leq t < \infty. \tag{2.44}$$

Therefore, our bpJ maximizing task has been reduced to maximizing the entropy rate $h(T_-)$ subject to Lagrange multiplier constraints on information decrement, $\mathbb{E} \log T_-$, and

energy constraint, $\mathbb{E}T_-$. It is known [37] that the entropy $h(T_-)$ is maximized subject to constraints on $\mathbb{E}\log T_-$ and $\mathbb{E}T_-$ when T_- is a gamma distribution with two parameters denoted by κ and b, i.e.,

$$f_{T_-}(t) = \left[\frac{b^\kappa t^{\kappa-1} e^{-bt}}{\Gamma(\kappa)}\right] u(t), \tag{2.45}$$

which is the bpJ-maximizing distribution of neuron j's ISI duration, T, sans its refractory period which always has duration Δ.

We emphasize that the result in Eq. (2.45) holds even when some of the weights are negative, provided $\mathbb{E}[W] > 0$. This allows the model to capture to some extent the phenomena of inhibition and leakage, both of which are extant in the real scenario, thereby enhancing the model's breadth of applicability.

2.7 From the Integral Equation to the Differential Equation

The integral equation (2.46) below relates the following three quantities:

1. The to-be-optimized pdf $f_\Lambda(\lambda)$ of the arithmetic mean of the net afferent excitation and inhibition of neuron j during an interspike interval (ISI).

2. The conditional pdf $f_{T_-|\Lambda}(t|\lambda)$ of j's encoding of Λ into the duration T of said ISI.

3. The long-term pdf $f_{T_-}(t)$ of j's ISI durations.

$$\int_0^\infty d\lambda f_\Lambda(\lambda) f_{T_-|\Lambda}(t|\lambda) = f_{T_-}(t), \ \forall\ 0 \le t < \infty. \tag{2.46}$$

Moreover, we have

$$f_{T_-|\Lambda}(t|\lambda) = \sum_{m=1}^{\infty} P_m \cdot \frac{\lambda^m t^{m-1} e^{-\lambda t}}{(m-1)!}. \tag{2.47}$$

That is, the conditional pdf in question is a mixture of gamma pdf's with integer shape parameters $m = 1, 2, \ldots$ and common scale parameter λ. The m^{th} weight in this mixture is

the probability that exactly m excitatory/inhibitory synaptic events occur during the ISI, namely

$$P_m = P(W_1 + W_2 + \cdots + W_{m-1} < \theta \text{ and } W_1 + W_2 + \cdots + W_m \geq \theta), \tag{2.48}$$

where $W_1, W_2, \ldots, W_m$ are i.i.d. with cumulative distribution function (cdf) $F_W(\cdot)$. We have derived formulas for P_m for some reasonable choices of the pdf $f_W(\cdot)$ in Section 2.5. Let's first consider a case in which P_m is nonzero only for two consecutive values of m, say n and $n+1$ with respective probabilities $P_n = p$ and $P_{n+1} = 1 - p := q$.[4] In such a case

$$\int_0^\infty d\lambda f_\Lambda(\lambda) \left(\frac{p\lambda^n t^{n-1}}{(n-1)!} + \frac{q\lambda^{n+1} t^n}{n!} \right) e^{-\lambda t} = f_{T_-}(t). \tag{2.49}$$

In [5] it was shown that, if $f_\Lambda(\cdot)$ is chosen so that bits communicated per joule expended is maximized, then $f_{T_-}(\cdot)$ must be a gamma pdf, i.e., there must exist parameters $\kappa > 0$ and $b > 0$ such that

$$f_{T_-}(t) = \left[\frac{b^\kappa t^{\kappa-1} e^{-bt}}{\Gamma(\kappa)} \right] u(t). \tag{2.50}$$

In Section 2.6 we showed that, even in cases with unequal synaptic weights including both excitatory and inhibitory synapses, the $f_\Lambda(\cdot)$ that maximizes bits conveyed per joule expended still must induce a Gamma-distributed T_-; i.e., Eq. (2.50) continues to be satisfied. When (i) Eq. (2.50) is substituted into the right-hand side of Eq. (2.49), (ii) t is changed throughout to s, and (iii) both sides are multiplied by $(n-1)!s^{1-n}$, the result is

$$\int_0^\infty d\lambda f_\Lambda(\lambda) \lambda^n \left(p + \frac{q\lambda s}{n} \right) e^{-\lambda s} = (n-1)! \cdot \frac{b^\kappa s^{\kappa-n} e^{-bs}}{\Gamma(\kappa)} \ . \tag{2.51}$$

[4] This occurs, for example, when $\theta = 19$ and each W is a binary r.v. that equals 6 with probability p_6 and equals 8 with probability $p_8 = 1 - p_6$. Then $P_4 = p_6^3$, $P_3 = 1 - p_6^3$, and all other P_n's are zero; hence, $m = 3$ and $q = p_6^3$.

Next, we use the fact that

$$\frac{q\lambda s}{n}e^{-\lambda s} = -\frac{q\lambda}{n}\cdot\frac{d(e^{-\lambda s})}{d\lambda}, \tag{2.52}$$

which we shall refer to as "the q-term." The portion of the left hand side of Eq. (2.51) that corresponds to the q-term is

$$-\int_0^\infty d\lambda f_\Lambda(\lambda)\lambda^{n+1}\frac{q}{n}\cdot\frac{d(e^{-\lambda s})}{d\lambda}. \tag{2.53}$$

Since the $d\lambda$ in the numerator of the integrand in Eq. (2.53) cancels that in the denominator, Eq. (2.53) is of the form $-\int UdV$ where

$$U(\lambda) = \lambda^{n+1}f_\Lambda(\lambda)\frac{q}{n} \tag{2.54}$$

and

$$V(\lambda) = e^{-\lambda s}. \tag{2.55}$$

This permits us to evaluate the q-term by means of the formula for integration by parts, namely $-\int UdV = -UV + \int VdU$. Note that for U and V as above, we have $U(0) = 0$ and $V(\infty) = 0$ so $U(0)V(0)$ and $U(\infty)V(\infty)$ both equal zero. It follows that the q-term equals

$$\int_0^\infty VdU = \int_0^\infty e^{-\lambda s}\left[(n+1)\lambda^n f_\Lambda(\lambda)\frac{q}{n} + \lambda^{n+1}f'_\Lambda(\lambda)\frac{q}{n}\right]d\lambda. \tag{2.56}$$

Inserting this representation of the q-term into Eq. (2.51) yields

$$\int_0^\infty d\lambda\left[\lambda^n f_\Lambda(\lambda)\left(p + (n+1)\frac{q}{n}\right) + \lambda^{n+1}f'_\Lambda(\lambda)\frac{q}{n}\right]e^{-\lambda s} = (n-1)!\cdot\frac{b^\kappa s^{\kappa-n}e^{-bs}}{\Gamma(\kappa)}\ . \tag{2.57}$$

Observe that the left hand side of Eq. (2.57) is the Laplace transform from the λ-domain into the s-domain of the function of λ that appears within the square brackets. It follows

that, if we apply the inverse Laplace transform operator, $\mathcal{L}^{-1}$, to both sides of Eq. (2.57), the left hand side of the result will be the function of λ that currently resides within that square bracket, so

$$\lambda^n f_\Lambda(\lambda)\left(p+(n+1)\frac{q}{n}\right)+\lambda^{n+1}f'_\Lambda(\lambda)\frac{q}{n}=\mathcal{L}^{-1}\left\{(n-1)!\cdot\frac{b^\kappa s^{\kappa-n}e^{-bs}}{\Gamma(\kappa)}\right\}. \tag{2.58}$$

One can readily verify that the Laplace transform of $\frac{\Gamma(n)}{\Gamma(\kappa)\Gamma(n-\kappa)}(\lambda-b)^{n-\kappa-1}u(\lambda-b)$ is precisely the function being operated upon by $\mathcal{L}^{-1}$ on the right hand side of Eq. (2.58). It follows from these observations and from the identity $p+(n+1)\frac{q}{n}=1+\frac{q}{n}$ that

$$f_\Lambda(\lambda)\left(1+\frac{q}{n}\right)+\lambda f'_\Lambda(\lambda)\frac{q}{n}=\lambda^{-n}\frac{\Gamma(n)}{\Gamma(\kappa)\Gamma(n-\kappa)}(\lambda-b)^{n-\kappa-1}u(\lambda-b). \tag{2.59}$$

Eq. (2.59) is a powerful result in that it constitutes a conversion of the integral equation (2.49) for the maximum-bpJ excitation/inhibition intensity $f_\Lambda(\lambda)$, $\forall\ b\le\lambda<\infty$, into a first order linear differential equation for this same function. The differential equation has a λ-varying coefficient on its f' term - namely, a constant times λ itself. Nonetheless, it has an explicit analytical solution because there exists a general solution to any inhomogeneous first order linear differential equation with variable coefficients. In light of the above it is reasonable to conjecture that if there are exactly k consecutive values of m for which $P_m>0$, then the corresponding integral equation for $f_\Lambda(\cdot)$ can be converted to a linear differential equation of order $k-1$ with variable coefficients. In general, for k consecutive values of m for positive P_m's, Eq. (2.51) will become

$$\int_0^\infty d\lambda f_\Lambda(\lambda)\left(\sum_{j=0}^{k-1}\frac{P_{j+1}\lambda^{j+1}s^j}{\Gamma(j+1)}\right)e^{-\lambda s}=(n-1)!\cdot\frac{b^\kappa s^{\kappa-n}e^{-bs}}{\Gamma(\kappa)}\ . \tag{2.60}$$

After applying integration by parts and Laplace transform repeatedly, Eq.(2.60) finally turns out to be an inhomogeneous Cauchy-Euler equation with variable coefficients as below

$$\sum_{j=0}^{k-1}\sum_{i=0}^{j} P_{j+1}\frac{(j+1)!}{(j-i+1)!}\frac{1}{i!(j-i)!}\lambda^{(j-i)}f_{\Lambda}^{(j-i)}(\lambda) = \frac{b^{\kappa}}{\Gamma(\kappa)\Gamma(1-\kappa)}\lambda^{-1}(\lambda-b)^{-\kappa}u(\lambda-b). \tag{2.61}$$

Eq.(2.61) also has an analytical closed-form solution, which serves as the bpJ-maximizing pdf of neuron j's averaged afferent excitation/inhibition intensity Λ.

2.8 Summary

We have shown that, when neuron j is designed to maximize bits conveyed per joule expended, even though j's synapses no longer are being required to all have the same weight or be excitatory only, the pdf of the ISI durations continues to be a delayed gamma distribution as it was in [5] wherein all the weights were assumed to be equal. This happens despite the fact that the conditional distribution for T given Λ is now a mixture of gamma distributions instead of the pure gamma distribution that characterizes the special case of equal weights.

Additionally, we have implicitly determined the optimal distribution $f_{\Lambda}(\lambda)$ that characterizes the afferent excitation/inhibition intensity by (1) maximizing the Shannon mutual information rate given a constraint on the total energy cost that a cortical neuron expends for metabolism, postsynaptic potential accumulation, and action potential generation and propagation during one ISI; (2) converting the integral equation to a differential equation with a closed-form solution.

The energy efficiency of the human brain in terms of information processing is astonishingly superior to that of man-made machines. By extending the information-energy efficient neuron model to a more general framework including both unequal synaptic weights and inhibitory synapses, the theory comes into closer correspondence with the actual neurophysiology of

cortical networks, which might pave the way to wider applications in neuroscience and engineering.

Chapter 3

Theoretical and Computational Neuron Modeling: Gaussian Diffusion Model

> *Scientific discovery consists in the interpretation for our own convenience of a system of existence which has been made with no eye to our convenience at all. One of the chief duties of a mathematician in acting as an advisor to scientists is to discourage them from expecting too much of mathematicians.*
>
> *-Norbert Wiener*

We develop a linkage between the mathematical analysis of a single neuron and the statistical connection of that neuron to the rest of the brain. The core of a stochastic neuron model is the selection of a conditional probability density, $f_{T|\Lambda}(\cdot|\lambda)$, for the random time T that it takes the neuron's postsynaptic potential to cross a possibly varying threshold given that the neuron's random excitation intensity Λ has assumed a particular value λ. For reasons we develop in detail, we have selected a certain subfamily of inverse Gaussian (IG) probability densities to serve in this capacity. We assume the neuron is energy efficient in the sense that it maximizes the Shannon mutual information it conveys to its targets per Joule of energy

it expends to generate and propagate its train of neural spikes. Using information theory, calculus of variations and Laplace transforms, we derive and solve a pair of coupled integral equations that describe how Λ must be distributed in order for the neuron to maximize bits transmitted per Joule expended (bpJ). The first equation's solution establishes that the at-this-point unknown bpJ-maximizing probability density $f_\Lambda(\lambda)$ must induce via $f_{T|\Lambda}(\cdot|\lambda)$ a random ISI duration whose probability density $f_T(t)$ belongs to the generalized inverse Gaussian (GIG) family. The algebraic shape factor of this $f_T(t)$ has the form $t^{-(\frac{3}{2}+D)}$ where $D > 0$, as compared with the standard IG density's shape factor $t^{-\frac{3}{2}}$. This result agrees with work on best matching of experimentally observed ISI durations reported in the literature. The solution of the second integral equation yields the exact form of the bpJ-maximizing $f_\Lambda(\lambda)$. This formula for $f_\Lambda(\lambda)$ is our principal result in that Λ is created not by the neuron being modeled but by those of the brain's neurons whose spike trains are afferent to one or more of the modeled neuron's excitatory synapses. Accordingly, $f_\Lambda(\lambda)$ serves as the abovementioned bridge that specifies how an energy efficient brain needs to match the long-term statistics of each of its neuron's inputs to that neuron's particular design.

3.1 Problem Statement and Preliminaries

3.1.1 Neuron Outputs and Inputs

As in [5] [6] [7] we call the neuron under study "neuron j", or just j for short. The output of j is a train of effectively identically shaped narrow voltage spikes, also known as action potentials (AP's). The durations of the interspike intervals (ISI's) between j's AP's constitute a random sequence which we denote by $\{T_k, k = 0, 1, 2, \dots\}$.

Neuron j's external excitation/inhibition comprises a collection of neural spike trains generated by the afferent presynaptic neurons, called j's afferent cohort. j's afferent synapses are connections between these neurons and j by means of which each such spike train becomes afferent to j. When a presynaptic spike arrives at one of j's presynaptic terminals, the

depolarization it provides opens calcium channels in the presynaptic bouton. A vesicle containing neurotransmitters fuses with the presynaptic membrane and releases an amount of neurotransmitters based on the size of the vesicle. Those neurotransmitters diffuse within the synaptic cleft. Some escape, but most bind to the postsynaptic receptors, which will cause the opening of the ion channels resulting in a net gain of positive/negative voltage across the postsynaptic membrane, an EPSP/IPSP. Each time the membrane potential crosses a possibly time-varying threshold value, j generates another output spike that propagates along its axon to the efferent neurons, referred to as j's efferent cohort or targets. A typical neuron in primate sensory cortex possesses circa $10,000$ inputs and circa $10,000$ outputs.

Each time j emits an efferent spike, there ensues a refractory period during which j cannot produce another spike. Neuroscientists distinguish two types of refractory periods - absolute and relative. Limiting attention to absolute refractoriness, by far the most common type under natural conditions, permits us to assume with negligible error that j's refractory periods all have the same duration, call it Δ. During each refractory period j strives to replenish supplies of chemicals in various of its compartments in anticipation of producing a subsequent spike in response to the ongoing afferent excitation. However, j's state (i.e., the chemical concentrations in all its subcompartments) does not always return to the same value at the end of each refractory period. In particular, in the not infrequent case in which j experiences several successive ISI's whose durations are only slightly larger than Δ, diminished supplies may prevent full restoration of the concentration of certain chemical species in some of j's subcompartments before the refractory period ends.

3.1.2 The Definition of Λ

The arrival of AP's at the presynaptic terminals of j's afferent excitatory synapses, often referred to as j's "bombardment," can be measured in units of spikes/s. Neuroscientists sometimes consider this arrival to be j's afferent excitatory intensity, usually denoted by λ; indeed, we did this in our earlier work [5] [6] [7]. Here, however, we prefer to define λ

at a finer level of electrochemical arrivals. One way of doing this is to define it to be the sum over all of j's excitatory synapses of the rate at which neurotransmitter molecules are traversing from the presynaptic to the postsynaptic side and binding to receptors located on j. A further refined definition would be that λ is the net algebraic rate at which ionic charges are accruing on j's cell membrane via diffusion and thereby are contributing to the PSP accumulation, taking into account that some of the charges deposited subsequently leak off through conductances to ground. Although this further refined definition has the perhaps disadvantageous property of not being wholly afferent in spirit, it corresponds to the effective λ that j itself is capable of measuring since. That's because, as an ISI progresses, its PSP accumulation is j's only summative response to the spikes afferent to its afferent excitatory synapses. Accordingly, we shall assume that at every time instant t there is an instantaneous random intensity at which contributions to j's PSP accumulation are arriving, which intensity we denote by $\Lambda(t)$. Once these contributions accrue to the extent that the PSP accumulation at j's axon hillock reaches a possibly time-varying threshold, j generates a new AP, propagates it along its axon to all its targets, and simultaneously begins a new ISI the first Δ seconds of which constitute a refractory period.

In order for the PSP accumulation to reach threshold, j must assemble a huge number of ions from a plethora of afferent postsynapses that are bombarded rather asynchronously and are located at widely differing relative distances from the axon hillock. Accordingly, Stein-Chen theory [38] implies that the random times at which said ionic contributions accrue are well-modeled by an inhomogeneous Poisson the intensity function of which is a realization $\{\lambda(t)\}$ of the aforementioned random process $\{\Lambda(t)\}$. That the statistics of $\{\Lambda(t)\}$ remain time-inhomogeneous is essential because j's *raison d'etre* is to report to all its targets about how the realization $\{\lambda(t)\}$ unfolds versus time. The highly refined definition of λ that we have adopted herein obviates the need to be concerned with the probability distributions governing such quantities as the weights of j's synapses or the number of molecules of neurotransmitter that reside in a randomly chosen presynaptic vesicle. With the basic building block of

excitation intensity set at the ion level, distributions such as these already have been averaged over in the process of defining $\{\Lambda(t)\}$.

3.2 Mutual Information and Energy Consumption

The astonishing capabilities of mammalian brains are in large measure attributable to their exceptional energy efficiency. Most neurons expend energy at an average rate that lies between half a nanowatt and two nanowatts. Since the human brain contains circa 3×10^{10} neurons, the whole brain expends less energy per second than does a 100 watt light bulb, the current estimate being circa 20 watts. [1] Since neurons are expressly designed to exchange information with one another, it has long seemed to us [5] [6] [7] [16] [17] [39] that each neuron's design should maximize the ratio of the rate at which it conveys information to the rate at which it expends energy. Hence, we shall impose below the performance criterion that a neuron addresses its tasks in a fashion that maximizes the bits it conveys to its efferent cohort per Joule it expends to calculate and propagate said bits. Simply put, we assume that neurons maximize bits per joule (bpJ). In this regard Laughlin and Sejnowski report in their survey article [27] that the belief that neuronal networks are constructed to be maximally energy efficient is attracting an increasing number of adherents in the neuroscience community. In what follows it will be seen that imposing a max-bpJ condition on single neuron function not only allows us to obtain key analytical conclusions that are in good agreement with experimental observations but also yields an intriguing bridge between single neuron theory and the theory of real neural networks.

[1] Recently, a team of neuroscientists, mathematicians, electrical engineers, and computer scientists (among others) generated a successful simulation of some 10 million neurons in cat visual cortex. The connectivity matrix of the neurons was precise in that it was based on painstaking tracking of the neurons' axons through thousands of tomographic sections of a real cat brain. Powerful computers were parallel programmed such that the simulation processed its inputs as fast as a cat does, albeit arguably (and it was fiercely and publicly argued) not to the same effect achievable by a real cat. Upon conducting an energy audit of this simulation, we estimated that it consumed some 10^8 times as many Joules per neuron than would the cat. Emailing our estimate to the project leader, Dharmendra Modha of IBM Almaden, brought his crisp and forthright reply, "No, 10^9!"

3.2.1 Information Rate

The mathematical instantiation of neuron j's rate of information production is Shannon's average mutual information rate between j's random excitation $\{\Lambda(t)\}$ and j's output $\{T_k\}$. Although information theory has been generalized to deal with abstract random objects such as the uncountable infinity of random variables that constitute $\{\Lambda(t)\}$ [40], it is highly challenging to evaluate such abstractions. Accordingly, as we did in [5] [6] [7], we approximate $\{\Lambda(t)\}$ by a random sequence $\{\Lambda_k\}$ defined as follows. Choose $t = 0$ to correspond to an instant randomly chosen from all those at which j has fired a spike. Then let T_1 denote the random time at which j fires its next spike, $T_1 + T_2$ denote the least time strictly greater than T_1 at which j fires another spike, and so on. Thus, T_k is the duration of j's k^{th} ISI henceforth called ISI_k. For each $k \in \{1, 2, \dots\}$, define

$$\Lambda_k = \frac{1}{T_k - \Delta} \int_{S_{k-1}+\Delta}^{S_k} \Lambda(t)dt, \tag{3.1}$$

where $S_0 = 0$ and for each $k \geq 1$, $S_k = T_1 + T_2 + \cdots + T_k$. In words, throughout the post-refractory portion of ISI_k we replace the random process $\Lambda(t)$ by a constant intensity Λ_k that equals the time average of $\Lambda(t)$ over that portion of ISI_k. For further discussion of and partial justification for this so-called Mean Value Assumption, the reader is referred to [5]. Using it, the mutual information rate between j's input and output, measured in bits/ISI, becomes

$$I := \lim_{n\to\infty} n^{-1} I(\Lambda_1, \Lambda_2, \dots, \Lambda_n; T_1, T_2, \dots, T_n), \tag{3.2}$$

where $I(X;Y)$ equals Shannon's mutual information between the random objects X and Y. (See any of [32], [33], [34], [40], or [41].) In [5] [7] we showed that, with considerable generality, the dominant behavior of the limit in Eq.(3.2) is the average mutual information between a single r.v. Λ_k and its associated r.v. T_k. Since our initial ISI has been chosen purely at random from all of j's ISI's, $I(\Lambda_k; T_k)$ does not depend on k. Therefore, we write simply $I(\Lambda; T)$, where T is the random duration of an ISI and the random variable Λ is the time

average of $\Lambda(t)$ over the post-refractory portion of that ISI.

3.2.2 Energy Expenditure

We now identify five energy expenditures $g_i(\lambda, t), 1 \leq i \leq 5$, that neuron j incurs during an ISI in which the random 2-vector $(\Lambda, T - \Delta)$ assumes the value (λ, t). Later, we associate these g_i's with terms in the exponent of our stochastically modeled neuron channel $f_{T|\Lambda}(t|\lambda)$.

- $g_1(\lambda, t) = A$, a positive constant. An example of this is the energy j must expend in order to propagate the neural spike it generates at the end of the ISI to all its target neurons. This and other fixed costs per ISI sum to A.

- $g_2(\lambda, t) = Bt$, where B is a positive constant. An example of this is the metabolic energy the neuron expends keeping itself healthy and nourished during an ISI. This grows linearly with the duration of the ISI whether j is processing relatively intensely during it or not.

- $g_3(\lambda, t) = C\lambda t$, C a positive constant. This term addresses the energy j devotes to generating contributions to the PSP accumulation by processing all the spikes afferent to it during the ISI. [2]

- $g_4(\lambda, t) = G/t$, G being a positive constant. g_4 charges a high energy costs for small values, t, of the random variable $T - \Delta$. That such a small-t penalty must exist is apparent from the definition of Δ as the least amount of time after one AP that another can be produced in response to natural afferent excitation. Although an experimenter can evoke AP's separated by less than Δ by injecting a large current into the axon hillock, this requires supplying power far exceeding what the neuron can expend on

[2] In order for $C\lambda t$ not to be an oversimplification, there needs to be, regardless of the value of λ, enough separate contributions to PSP during an ISI that the coefficient of variation (standard deviation divided by mean) of the number of them and that of the weights of the synapses that generate them both are smaller than, say, 0.1. Primary cortical neurons, among others, conform to these conditions.

its own. It stands to reason that the closer to $t = 0$ the neuron produces an AP in response to natural stimulation, the more energy it takes to do so. [3]

- $g_5(\lambda, t) = D \log t$. If $D > 0$, then this logarithmic term charges an energy penalty whenever $t > 1$ and "rewards" t-values smaller than the time unit, and conversely, if $D < 0$. Note that this logarithmic term is dominated by $g_2 = Bt$ when t is large and by $g_4 = G/t$ when t is small, provided D is not many times larger than either B or G, respectively. Hence g_5 serves as a fine tuning adjustment to g_2 and/or g_4. It is widely felt that most real neurons are finely tuned. In Section 3.6 we find it necessary to assume that $D > 0$ in order to get closed-form analytical results.

3.3 The GIG and IG Probability Densities

A nonnegative random variable (r.v.) that has a probability density of the form

$$f(t) = Ct^{\alpha-1} \exp\left(-\gamma t - \frac{\beta}{t} \right), 0 < t < \infty;\ \beta > 0, \gamma > 0, \tag{3.3}$$

is an instance of a generalized inverse Gaussian r.v., henceforth a GIG r.v. The value of the reciprocal of the normalizing constant C is (see [36], Eq. 3.471.9) $C^{-1} = 2(\beta/\gamma)^{\frac{\alpha}{2}} K_\alpha(2\sqrt{\beta\gamma})$, where $K_\alpha(\cdot)$ is the modified Bessel function of the second kind with index α. The GIG pdf generalizes the key special case $\alpha = -\frac{1}{2}$, known as the inverse Gaussian (IG) pdf, in which the power of t is $\alpha - 1 = -\frac{1}{2} - 1 = -\frac{3}{2}$ and $C^{-1} = \sqrt{\pi/\beta}\exp(-2\sqrt{\beta\gamma})$, where we have used the fact (see [36], Eq. 8.469.3) that $K_{\pm\frac{1}{2}}(z) = \sqrt{\pi/(2z)}e^{-z}$. Without loss of generality, we henceforth set $\gamma = 1$, this being tantamount to selection of the t-unit.

[3] At present, the neuroscience community's understanding of the functional form of the t-dependence of this energy expenditure is insufficient to guarantee that it is proportional to the reciprocal of t as opposed to, say, some other negative power of t. Therefore, our use of G/t, though mathematically compelling in the sequel, amounts to a convenient educated guess.

3.3.1 IG and GIG in the Neuroscience Literature

Several investigators have used the GIG pdf to model a neuron viewed as a communication channel, the input of which is what we have called Λ and the output of which is what we have called T. Perhaps the first to do so were Gerstein and Mandelbrot [42]. Barndorff-Nielsen et al., put the GIG pdf on a firm mathematical foundation by showing it characterizes the random time it takes certain diffusions with drift to hit a fixed-height boundary [43] [44]. Specifically, GIG pdf's with $\alpha < -1/2$ (respectively, $\alpha > -1/2$) are associated with models in which the boundary attracts (repels) the diffusing entity. The IG case $\alpha = -1/2$ yields a neutral boundary that neither attracts nor repels. Recently, some statisticians and neuroscientists have argued that GIG pdf's provide better fits to the empirical distribution of ISI durations than do gamma pdf's and lognormal pdf's that often have been so used [45] [46].

3.3.2 The IG Neuron Channel Model

In a classic paper Schrödinger [47] showed that the time it takes what now is generally called a Wiener process with negative drift to fall from an initial height to zero is IG distributed. Wiener processes have infinitesimal Gaussian increments, whereas we have argued that the process we call $\{\Lambda(t)\}$ has fine-grained positive increments that occur in accordance with an inhomogeneous Poisson process. Within a given ISI, however, we have agreed in Section 3.2 to adopt a homogeneous Poisson model with a random intensity Λ that equals the arithmetic average of $\Lambda(t)$ over that ISI. Under this Mean Value Assumption the PSP accumulation behaves as a homogeneous filtered Poisson process during said ISI the increments of which arrive at a fixed rate λ that is a realization of the r.v. Λ. These increments are not Gaussian, however. Rather, they are a multitude of ionic charges with a positive mean, denoted by m. [4] Also, their standard deviations are small because most of the ions have a charge of magnitude 1 or 2, with $^{+}$ more likely than $^{-}$ and $^{++}$ more likely than $^{--}$. Since λt is of

[4] Some of the increments may be negative, corresponding to ion charges leaking off the PSP membrane to ground, but the mean amplitude is positive.

the order of 10^6 (likely more) at the time t at which threshold is crossed, we expect that the PSP accumulation then, call it $V(t)$, will be nearly Gaussian with mean $m\lambda t$ and variance near $m^2\lambda t$, hence standard deviation near $m\sqrt{\lambda t}$. It therefore has a coefficient of variation close to $1/\sqrt{\lambda t} \simeq 10^{-3}$, indicating high accuracy and hence capable information transfer. The third central moment of a Poisson r.v. with intensity ν is equal to ν. In this case under consideration $\nu \simeq m^3\lambda t$, with the proportionality constant positive and less than 1. This is because positive and negative displacements from the mean $m\lambda t$ have opposite signs when cubed and are close to evenly balanced with a slight edge toward positivity. The principal distinction between the actual pdf of $V(t)$ and the Gaussian approximation thereto that uses $V(t)$'s own mean and variance resides in this third central moment, since a Gaussian symmetry around the mean forces all odd-order central moments to be 0. In the current context, however, we see that the nonzero central moment shifts $V(t)$ for t near the threshold crossing time by a fractional amount that is $O(\sqrt[3]{m^3\lambda t})/(m\lambda t) = O((\lambda t)^{-2/3}) = O(10^{-4})$. This implies that the Gaussian approximation is almost perfect near its mean, which implies that we may safely use Schrödinger's IG distribution as our stochastic neuron channel model. Accordingly, in what follows we write

$$f_{T|\Lambda}(t|\lambda) = C_\lambda t^{-3/2} \exp\left(-\gamma\lambda t - \frac{\beta}{\lambda t}\right), \tag{3.4}$$

$2\gamma = (m/\sigma)^2$, $2\beta = (\theta/\sigma)^2$, m and σ^2 are the mean and variance of the charges of the ionic contributions to and leakages from the membrane potential, and θ is the AP threshold voltage. Again, we fix the time unit by setting $\gamma = 1$, whereupon we obtain the IG pdf

$$f_{T|\Lambda}(t|\lambda) = C_\lambda t^{-3/2} \exp\left[-\lambda t - \frac{(\theta/m)^2}{\lambda t}\right]. \tag{3.5}$$

3.3.3 Λ-Dependent Thresholds

We stress that θ may depend on the value λ upon which we are conditioning in Eq.(3.5). Despite decades of neuroscientific experimentation, little is known about the behavior of the threshold during an ISI. The membrane potential $V(t)$ can be measured, but there is no way of measuring how much above $V(t)$ the threshold lies at time t; the only time an experimenter knows the level of the threshold is the instant at which it is crossed. [5] However, in our ionic model $m\lambda(t)$ is the rate at which the PSP accumulation is increasing in the immediate vicinity of time t. This permits computation of the realization λ of the r.v. Λ at the moment the threshold is crossed as follows. At any elapsed time since j's last AP, call it s, the only summative pieces of information j has at its disposal about the net excitation it has received during $(0, s]$ are s itself (which is known everywhere throughout j) and $V(s)$, the current value of the membrane potential at the axon hillock from which j's AP's are emitted. These, however, suffice for j to produce an estimate of the arithmetic average of $\Lambda(r)$ over $0 < r \leq s$ that becomes increasingly accurate as s grows. j has several means at its disposal to adjust the AP threshold based on said estimate, among which are to genuinely raise or lower the voltage threshold and/or to adjust the conductance from the PSP membrane to ground to hasten or to delay the instant at which the threshold will be crossed. Although it may have appeared impossible for the threshold to depend upon λ because, as we have defined it, λ is not determined until the threshold is crossed, the preceding discussion reveals that it is well within j's ability to do effectively that very thing.

Let us assume for simplicity that the threshold's dependence on λ is of the form

$$\theta(\lambda) = c\lambda^{\mu}, \ c > 0. \tag{3.6}$$

For $\mu > 1$ the threshold eventually grows asymptotically faster in λ than does the linearly

[5] Also, the instantaneous excitatory intensity $\Lambda(t)$ is a non-observable abstraction, as is its arithmetic average Λ over the post-refractory portion of the ISI. Even realizations λ of the r.v. Λ are not directly observable at present because we have no way of simultaneously observing all the spike trains that are afferent to j's circa 10,000 postsynaptic terminals which are densely packed in 3D along its dendrites and soma.

growing expected value of the PSP, so before long there would be an ISI that never ends. Moreover, we cannot have $\mu < 0$ because that would crush the $T - \Delta$'s of all the ISI's into a small range just above 0, thereby precluding accurate estimation of λ based on t, which is inconsistent with our goal of energy efficient information transfer. Therefore, we need consider only $0 \leq \mu \leq 1$. $\mu = 0$ is the classical constant threshold for which we obtain

$$\mathbb{E}[T|\Lambda = \lambda] = \frac{\text{const.}}{\lambda}, \tag{3.7}$$

which has long and widely been felt to be the true nature of the inverse relationship between $\mathbb{E}T$ and λ. However, that inference is unjustified because it is based on the misconception that someone has physically measured λ, which we have noted is not a quantity that currently can be measured. The "λ" that appears in the denominator of Eq.(3.7) might be any of many monotonic functions of the "true" λ. In particular, if $\theta's$ dependence on the "true" λ is $b\lambda^{\mu}$ for some $\mu \in [0, 1)$, then $\mathbb{E}[T|\Lambda = \lambda]$ will be const./$\lambda^{1-\mu}$, hence monotonic decreasing in λ as required to be in conformance with neuroscientific observations.

We henceforth set $\mu = 1/2$, right in the center of the allowable range. The compelling motivation for this choice is not its centrality; rather, it's that when $\mu = 1/2$ we are able to solve exactly for the sought-after bpJ-maximizing $f_{\Lambda}(\lambda)$, as described below.

3.4 Mathematical Analysis of the IG Neuron Channel with $\mu = 1/2$

For any choice of $f_{\Lambda}(\lambda)$, the corresponding pdf of $T - \Delta$ is given by the integral equation

$$f_{T-\Delta}(t) = \int_0^{\infty} f_{T-\Delta|\Lambda}(t|\lambda) f_{\Lambda}(\lambda) d\lambda, \ t \geq 0. \tag{3.8}$$

From Appendix A of [5] [6] [7], we know that $f_\Lambda(\cdot)$ is bpJ-maximizing if and only if the $f_{T-\Delta}(\cdot)$ that it generates via equation (4.7) satisfies

$$\int_0^\infty dt f_{T-\Delta|\Lambda}(t|\lambda)\left[\log(\frac{f_{T-\Delta|\Lambda}(t|\lambda)}{f_{T-\Delta}(t)}) - \sum_i^m g_i(\lambda,t)\right] = 0 \;\forall\; \lambda \geq 0, \tag{3.9}$$

where the $g_i, 1 \leq i \leq 5$, are the energy terms itemized in Section 3.2.

3.4.1 Finding $f_{T-\Delta}(t)$ when $\alpha = -1/2$ and $\mu = 1/2$

With $\mu = 1/2$, the θ^2 in the numerator of the $1/t$ term in the exponent of Eq.(3.5) becomes proportional to λ, thereby canceling the λ in the denominator and reducing the coefficient of $1/t$ in the exponent to a positive constant which we denote by b. Setting $\alpha = -1/2$ to obtain an IG neuron channel and substituting the functional forms of the g_i's into Eq.(4.8) gives

$$\begin{aligned}&\int_0^\infty dt C_\lambda e^{-\lambda t - \frac{b}{t} - \frac{3}{2}\log t}\left[\log C_\lambda - \lambda t - \frac{b}{t} - \frac{3}{2}\log t - A - Bt - C\lambda t - D\log t - \frac{G}{t}\right]\\ &= \int_0^\infty dt C_\lambda e^{-\lambda t - \frac{b}{t} - \frac{3}{2}\log t}\log f_{T-\Delta}(t).\end{aligned} \tag{3.10}$$

Because the term $\log C_\lambda$ in the square bracket in Eq.(3.10) does not depend on t and $C_\lambda e^{-\lambda t - \frac{b}{t} - \frac{3}{2}\log t}$ is the conditional pdf $f_{T-\Delta|\Lambda}(t|\lambda)$, the left-hand side of Eq.(3.10) equals $\log C_\lambda$ plus a revised version of said left-hand side in which $\log C_\lambda$ no longer is present in the square bracket. Making that change and then dividing by C_λ yields the next equation in which we also have substituted s for λ throughout in anticipation of solving the equation via Laplace transforms.

$$\begin{aligned}&-A\int_0^\infty dt e^{-st}e^{-\frac{b}{t}}t^{-\frac{3}{2}} - (G+b)\int_0^\infty dt e^{-st}e^{-\frac{b}{t}}t^{-\frac{5}{2}} - B\int_0^\infty dt e^{-st}e^{-\frac{b}{t}}t^{-\frac{1}{2}}\\ &-(C+1)s\int_0^\infty dt e^{-st}e^{-\frac{b}{t}}t^{-\frac{1}{2}} - \left(\frac{3}{2}+D\right)\int_0^\infty dt e^{-st}e^{-\frac{b}{t}}t^{-\frac{3}{2}}\log t - C_s^{-1}\log C_s^{-1}\\ &= \int_0^\infty dt e^{-st}e^{-\frac{b}{t}}t^{-\frac{3}{2}}\log f_{T-\Delta}(t).\end{aligned} \tag{3.11}$$

Note that the right-hand side of Eq.(3.11) is the Laplace transform of $e^{-\frac{b}{t}}t^{-\frac{3}{2}}\log f_{T-\Delta}(t)$. Also, each term on the left-hand side in which the only place s appears is in a factor e^{-st} in the integrand of an integral over t from 0 to ∞ is the Laplace transform of the function of t that constitutes the other factor of said integrand. There are only two terms on the left-hand side that do not have that form, namely $-C_s^{-1}\log C_s^{-1}$, and $-(C+1)s\int_0^\infty dte^{-st}e^{-\frac{b}{t}}t^{-\frac{1}{2}}$. The latter of these is recognized as the Laplace transform of $-(C+1)\frac{d}{dt}(e^{-\frac{b}{t}}t^{-\frac{1}{2}})$. Therefore, applying the inverse Laplace transform operator $\mathcal{L}^{-1}$ to both sides of Eq.(3.11) yields

$$e^{-\frac{b}{t}}t^{-\frac{3}{2}}\left[-A-(G+b)t^{-1}-Bt-\left(\frac{3}{2}+D\right)\log t\right]-(C+1)\frac{d}{dt}(e^{-\frac{b}{t}}t^{-\frac{1}{2}})-\mathcal{L}^{-1}(C_s^{-1}\log C_s^{-1})$$
$$=e^{-\frac{b}{t}}t^{-\frac{3}{2}}\log f_{T-\Delta}(t). \tag{3.12}$$

It follows that, if we can evaluate the inverse Laplace transform from the s-domain to the t-domain of the function $-C_s^{-1}\log C_s^{-1}$ on the left-hand side of Eq.(3.12), then exponentiation of the resulting equation provides an explicit formula for $f_{T-\Delta}(t), 0\le t<\infty$, the pdf of the ISI duration (sans refractory period) that is induced by the bpJ-maximizing $f_\Lambda(\lambda)$. Once in possession of said $f_{T-\Delta}(t)$, we proceed in Section 3.5 to solve Eq.(4.7) for the bpJ-maximizing $f_\Lambda(\lambda)$.

3.5 Evaluating the Needed Inverse Laplace Transform

In the subfamily of IG distributions with $\gamma=1$ and $\mu=1/2$, we have

$$C_s^{-1}=2(b/s)^{-\frac{1}{4}}\sqrt{\pi/4\sqrt{bs}}e^{-2\sqrt{bs}}=\sqrt{\pi/b}e^{-2\sqrt{bs}}$$

and therefore

$$-C_s^{-1}\log C_s^{-1}=2\sqrt{\pi s}e^{-2\sqrt{bs}}-\sqrt{\pi/b}\log(\sqrt{\pi/b})e^{-2\sqrt{bs}}.$$

Fortunately, both $e^{-k\sqrt{s}}$ and $\sqrt{s}e^{-k\sqrt{s}}$, where k is a positive constant, have known inverse Laplace transforms given respectively by Eqs. (29.3.82) and (29.3.87) in [48]. When these are inserted and the algebra is carried through, we obtain

$$f_{T-\Delta}(t) = Kt^{-\frac{3}{2}-D}e^{-Bt-\frac{Q}{t}}, \tag{3.13}$$

where $K > 0$ is a normalizing constant, $Q = G + bC$, and both D and Q are positive. Note that $f_{T-\Delta}(t)$ is a GIG pdf with a shape factor exponent $-\frac{3}{2} - D$. Because $D > 0$, this GIG density's shape factor is more negative than that of an IG density. This agrees with fits to in vitro experimental data re ISI durations of certain goldfish retinal ganglion cells reported in [46], where each cell's ISI pdf was best fit by GIG pdf's as opposed to lognormals. Also, the α values were negative for all these cells and more negative than -1/2 for some of them. Given that these cells were not in behaving animals at the time and hence not receiving their normal excitation patterns over an extensive time interval, these results lend credence to the fact that goldfish RG cells may operate in the bpJ sensitive fashion we have postulated.

3.6 The Bits/Joule Maximizing $f_\Lambda(\lambda)$

In this section we find $f_\Lambda(\lambda)$, the bpJ-maximizing pdf of the time average of the afferent intensity over a randomly chosen ISI. We continue to assume that $\alpha = -1/2$ and $\mu = 1/2$ and to fix the time unit by setting $\gamma = 1$. The integral equation we need to solve for $f_\Lambda(\lambda)$ is

$$Kt^{-\frac{3}{2}-D}e^{-Bt-\frac{Q}{t}} = \int_0^\infty d\lambda f_\Lambda(\lambda) C_\lambda t^{-\frac{3}{2}} e^{-\frac{b}{t}} e^{-\lambda t}, \ \forall\ 0 \le t < \infty. \tag{3.14}$$

A factor of $t^{-\frac{3}{2}}$ is common to both sides of Eq. (3.14) so we cancel it. Multiplying both sides by $e^{\frac{b}{t}}$ and then temporarily replacing t by s throughout yields

$$\int_0^\infty d\lambda e^{-s\lambda} f_\Lambda(\lambda) C_\lambda = Ks^{-D} e^{-Bs} e^{-\frac{Q-b}{s}}. \tag{3.15}$$

The left-hand side of this equation is recognized as the Laplace transform from the λ-domain to the s-domain of the function $C_\lambda f_\Lambda(\lambda)$, so

$$C_\lambda f_\Lambda(\lambda) = K\mathcal{L}^{-1}_{s\to\lambda}\ (s^{-D}e^{-Bs}e^{-\frac{Q-b}{s}}). \tag{3.16}$$

Since all that a factor e^{-sB} contributes to an inverse Laplace transform is a delay by B, we may write

$$f_\Lambda(\lambda) = KC_\lambda^{-1}\cdot\mathcal{L}^{-1}_{s\to\lambda-B}\ (s^{-D}e^{\frac{b-Q}{s}}). \tag{3.17}$$

Determination of $f_\Lambda(\lambda)$ therefore has been reduced to finding the inverse Laplace transform of the function $s^{-D}e^{\frac{b-Q}{s}}$. Omitting unwieldy calculations in order to save space, we find with the help of Eqs. (29.3.81) and (29.3.80) of the Laplace transform table in [48] that the result we have been seeking is

$$f_\Lambda(\lambda) = KC_\lambda^{-1}(\frac{\lambda-B}{b-Q})^{\frac{D-1}{2}}I_{D-1}(2\sqrt{(b-Q)(\lambda-B)})u(\lambda-B), \tag{3.18}$$

where $K>0$ is a positive normalizing constant, $u(\cdot)$ is the unit step function and I_{D-1} is a modified Bessel function of the first kind with index $D-1$, which function is positive and integrable over its domain of definition for all $D>0$. This result echoes a finding in [5] [6] [7], wherein a more simplistic gamma pdf was used for $f_{T|\Lambda}(t|\lambda)$, that there is a range of low values of λ to which the bpJ-maximizing afferent intensity with which the rest of the brain excites an energy efficient neuron assigns zero probability. In the case of our IG channel model and with $\mu=1/2$, that interval is $0\leq\lambda\leq B$. A major contrast between our earlier gamma channel model and our current IG one is that all but a few low-order moments of the bpJ-maximizing $f_\Lambda(\lambda)$ were infinite in the gamma model, whereas all moments of the bpJ-maximizing $f_\Lambda(\lambda)$ given by Eq. (3.18) for the IG case with $\mu=1/2$ are finite.

Chapter 4

Theoretical and Computational Neuron Modeling: Non-Gaussian Diffusion Model

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

-Enrico Fermi

Neuronal information processing is energetically costly. Energy supply restrictions on information processing have caused brains to evolve to compute and communicate information with remarkable efficiency. Indeed, energy minimization subject to functional constraints is a unifying principle. Toward better comprehension of neuronal information processing and communication from an information-energy standpoint, we consider a continuous time continuous state-space neuron model with a generalized inverse Gaussian (GIG) conditional density. This GIG model arises from a Lévy diffusion process that is more general than that of a Wiener process with drift. We show that, when the GIG neuron operates so as to maximize bits per joule (bpJ), the output interspike interval (ISI) distribution is a related GIG marginal distribution. Also, we specify how to obtain the associated input distribution

$f_\Lambda(\lambda)$ numerically. Then we analyze how to maximize $I(\Lambda;T)$ per joule of energy the neuron expends when the energy cost function belongs to the family apropos of the GIG model. By generalizing from the Gamma and inverse Gaussian (IG) families to the GIG family, the derived results contain both the homogeneous Poisson and Wiener processes as special cases. The results allow us to readily compute the tradeoff between information rate (bits/second) and average power (Joules/second) in the GIG class, reminiscent of Shannon's celebrated formula for such curves for the additive Gaussian family.

4.1 Problem Statement and Preliminaries

4.1.1 Introduction

Neuronal information processing is energetically costly. The approximately 86 billion neurons composing the human brain, despite comprising only 2% of the body weight, consume on average 20% of the energy provided to the whole body when physically at rest [1] [2]. Energy supply restrictions on information processing have caused brains to evolve so as to compute and communicate information with remarkable efficiency. Indeed, energy minimization subject to functional constraints is a unifying principle [4]. There is reason to believe that neurons operate within a factor of 2 of the fundamental limit of $kT\log_2(e)$ Joules per binary write-read-erase cycle, k being Boltzmann's constant and, here only, T being body temperature in degrees Kelvin. Since neurons are expressly designed to exchange information with one another, We are led to believe that understanding the relationship between information processing and energy expenditure at the neuronal level is fundamental to understanding brain networks. This has led an increasing fraction of neuroscientists to subscribe to the view that each neuron's design should maximize the ratio of the rate at which it conveys information (in the sense of the aforementioned write-read-erase cycle) to the rate at which it expends energy. Hence we need to understand how an adequate supply of adenosine

triphosphate (ATP), the molecular unit of currency for energy transfer, is ensured for each neuron's information processing.

Information theory has often been applied in neuroscientific data analysis and biological systems modeling [12] [13] [15]. However, energy-efficient neural codes have only been studied for less than thirty years [16] [17] [18]. Evidence supporting energy efficiency has been reported for ion channel [4], action potential [19], synapse [20], photoreceptor [21], retina [22] [23], grey matter [24], white matter [25] and cortex [26]. Laughlin and Sejnowski discussed communication in cortical networks from an energy-efficiency point of view [27].

Toward better comprehension of neuronal information processing and communication from an information-energy standpoint, we earlier proposed and studied mathematical models of single neurons as engines of computation and communication based on both the homogeneous Poisson process [5] [6] and the Wiener process [8].

In this Chapter the goal is to extend the analysis to a model supported by a more general Lévy diffusion process. This results in a fixed threshold hitting time the probability density function (pdf) of which belongs to the three-parameter generalized inverse Gaussian (GIG) family. The GIG class subsumes many key two-parameter first passage time distribution families including the Gamma and inverse Gaussian (IG) distributions. By virtue of the infinitely divisible property of the GIG family, the associated GIG neuron channel model represents the random excitation and inhibition intensity at the axon initial segment in a manner that appeals both to neuroscientists and to mathematical statisticians, thereby promising to deepen our understanding of how information processing is performed at the ionic level, at the neuronal level, and perhaps eventually at the network level.

4.2 Neuronal Stimulation

A neuron in the brain receives stimulation in the form of individual spike trains from each member of a set of other neurons and/or sensory ganglia called the neuron's *afferent cohort*.

These spike trains each arrive at a unique synapse between the axon of a member of the afferent cohort and the dendrite tree or soma of the neuron we seek to analyze. We henceforth call that neuron "neuron j", or just "j" when that is unambiguous. In this Chapter we define the intensity of j's stimulation in a way that significantly differs mathematically from how we have in our earlier work. Neuronal stimulation usually is defined presynaptically by postulating a mathematical model of the totality of the afferent spike trains. Having to postulate a model is an unavoidable consequence of the following facts.

1. The average cardinality of the afferent cohorts of neurons in primary sensory cortex is circa $10,000$; certain neurons in the vestibular sensory modality (balance) are bombarded simultaneously by more than $100,000$ afferent spike trains.

2. It is not currently possible to make *in vivo* experimental measurements of the activity of all of these afferent spike trains at once; in fact, it has only recently become possible to measure the arrival of the spikes of even a single such train directly at the presynaptic terminal to which it is afferent.

3. Many synapses are excitatory and many others are inhibitory.

4. Controversy abounds as to whether spike trains from different members of the afferent cohort are effectively asynchronous or whether they often exhibit statistically significant synchrony. The argument against such synchrony is strong as regards the individual spike times in different afferent spike trains. On the other hand there is strong evidence of statistically significant correlation of spiking rates on the part of different members of the afferent cohort when said rates are calculated over time intervals comparable to the mean long-term interspike interval (ISI) duration.

Accordingly, to date and probably for some years or decades to come, analytical theorists have the combined advantage and challenge of characterizing afferent stimulation in a manner that is both physically reasonable and mathematically tractable.

The time at which j emits its next spike is determined not presynaptically but postsynaptically at j's axon hillock where j's "spike generator" is located. In response to the net afferent stimulation since the last time j generated a spike, a time-varying potential is produced on the postsynaptic cell membrane. The value of this membrane potential in the immediate vicinity of the spike generator is called the postsynaptic potential (PSP). When the PSP reaches a predetermined threshold value, j emits its next spike. Changes in the PSP over infinitesimally short intervals are effected by the depositing of ions onto the postsynaptic membrane and by the departing of ions that either leak to ground or are attracted elsewhere. Some of these ions are positively charged and others negatively. This suggests that despite the inherently presynaptic nature of j's afferent stimulation, its pragmatic effect resolves postsynaptically to the behavior versus time of the PSP comprising the diffusion of a multitude of ions. Modeling j's spiking behavior via this ionic diffusion has the desirable effects of allowing for both excitatory and inhibitory synapses and for PSP leakage, two crucial phenomena that are difficult to incorporate when one models j's stimulation presynpatically. Accordingly, we model the interspike interval (ISI) random variable T as the time it takes a certain diffusion of ionic particles that possesses an average positive drift rate of charge during the ISI to reach the spiking threshold. Although said average drift rate, which we denote by the random variable, Λ, must be positive in order to reach a positive threshold starting from an initial level we define to be 0, the PSP will exhibit both positive and negative fluctuations during the ISI.

4.2.1 GIG Hitting Times

The work of Barndorff-Nielsen et al. [43] [44] describes a time-homogeneous Markovian diffusion that, for a given value λ of a positive drift rate random variable (r.v.) Λ, creates a random first passage time (hitting time) to a fixed threshold level that possesses the GIG conditional pdf

$$f_{\tilde{T}|\Lambda}(t|\lambda) = C_\alpha \lambda^\alpha t^{\alpha-1} \exp\left(-\gamma\lambda t - \frac{\beta}{\lambda t}\right), \tag{4.1}$$

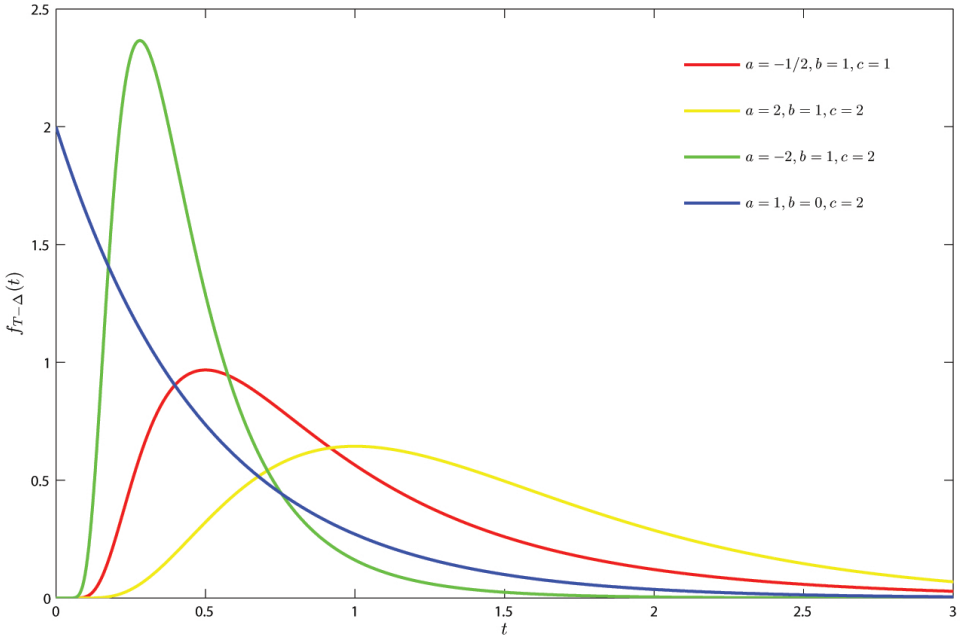

Figure 4.1: GIG densities for four choices of parameters (a, b, c) in Eq. (4.10).

where the normalizing constant's reciprocal is

$$C_\alpha^{-1} = 2\left(\frac{\beta}{\gamma}\right)^{\frac{\alpha}{2}} K_\alpha(2\sqrt{\beta\gamma}). \tag{4.2}$$

Here α is real, β is real and nonnegative, γ is real and positive, $K_\alpha(\cdot)$ is a modified Bessel function of the second kind of order α and $\tilde{T} := T - \Delta$, where Δ is the duration of the modeled neuron's refractory period. $f_{\tilde{T}|\Lambda}(t|\lambda)$ is called a GIG conditional pdf with parameters α, β and γ. Setting $\lambda = 1$ in Eq. (4.1) yields the unconditional GIG pdf several examples of which are shown in Figure 2.1.

In 1915 Schrödinger [47] considered a particle released at time 0 at height h above the ground whose altitude diffuses according to a Brownian motion in the earth's gravitational field, taking into account that in the steady-state the density of particles increases as one approaches the ground. Such a particle's mean negative drift rate is constant. If the magnitude of this drift rate is denoted by λ, then the pdf of the time at which the particle first hits the ground is given by the special case of Eq. (4.1) in which $\alpha = -1/2$, with β proportional to

h^2. The $\alpha = -1/2$ case is known as the IG pdf. In the context of neuronal PSP, the IG pdf governs the time it takes the PSP to hit a fixed positive threshold level θ if it starts at zero at the end of the neuron's refractory period and diffuses according to what modern parlance calls a Wiener process with positive drift rate λ; β is proportional to θ^2. That model of neuronal spiking, first introduced by Gerstein and Mandelbrot [42], clearly is the drift-reversed version of Schrödinger's problem with θ in the role of h. It is possible to extend the IG pdf to cases in which the threshold varies with the time t and/or is adapted in accordance with a time-varying estimate of the empirical drift rate that equals t^{-1} times the PSP level at time t [49]. We have several good reasons to generalize from the IG model to the GIG model. Here are three of them.

- The IG model lacks certain closure properties. The most crucial of these for our purposes is that, if the conditional pdf of $\tilde{T}$ given $\Lambda = \lambda$ is IG and the neuron is designed to maximize bits transmitted per Joule expended (bpJ), then the marginal pdf of $\tilde{T}$ will be GIG. This is a consequence of recent work of ours in which we have shown that, if the conditional pdf of $\tilde{T}$ given $\Lambda = \lambda$ is GIG(α, β, γ) and bpJ maximization prevails, then the marginal of $\tilde{T}$ is GIG(a, b, c). The parameter vector (a, b, c) differs from (α, β, γ). So, closure holds in this sense for the GIG family but not for the IG family (see Section 4.4).

- Allowing α to vary is crucial to achieving the rich variety of pdf shapes as partially illustrated in Figure 4.1. This is highly important to our current view that neuron j can adapt its parameters (α, β, γ), albeit rather slowly and over limited ranges. We conjecture in Section 4.3 that this plays a critical role in how j maintains bpJ maximization despite slow but continual changes over time in the probability distribution of the environmental stimulation that j receives from its afferent coherent.

- The strength of a model increases with the number n of its parameters for $n \leq 3$, but it decreases for $n > 3$. To bolster this arguable claim, we both refer and defer to the

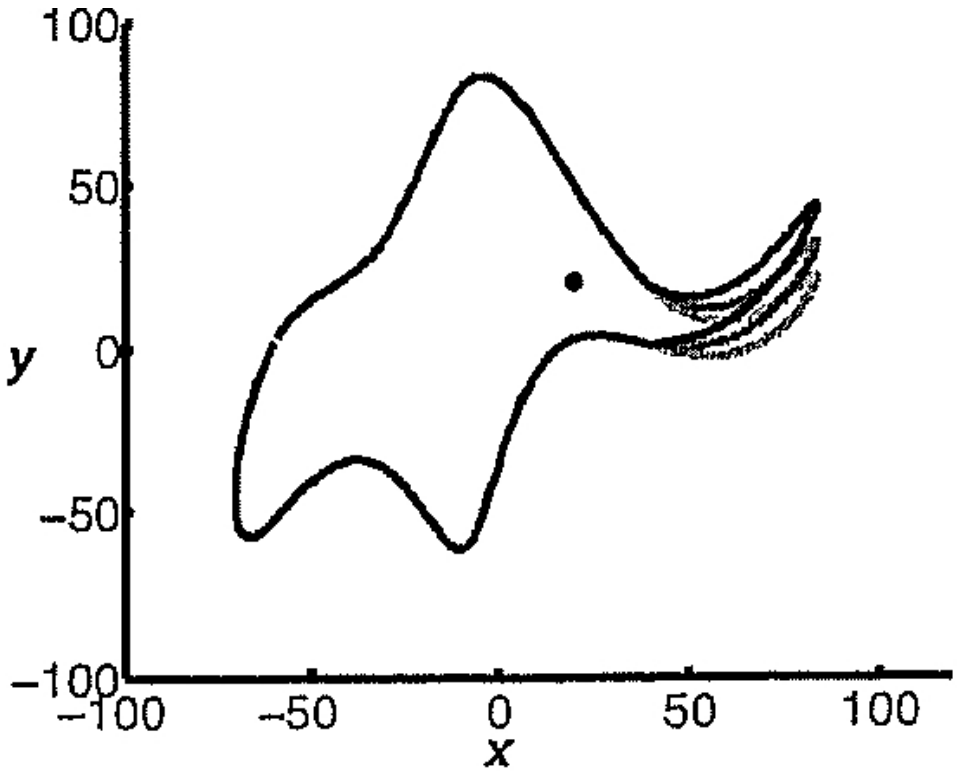

Figure 4.2: Outline of an elephant with the wiggling trunk. [51]

(in)famous remark of John von Neumann re "overmodeling" as quoted to Freeman Dyson by Enrico Fermi [50]:

> "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Although subsequent attempts at fitting an elephant needed dozens of parameters, a good 4-parameter fit was achieved in 2009 [51]; see Figure 4.2. However, the 4 parameters are complex, so this fit really uses 8 parameters; adding a ninth supports wiggling the trunk.

4.2.2 Modeling Neural Inputs and Outputs

We first introduce a mathematical framework for how our single neuron, j, stochastically processes and communicates information. The output of j is a temporal sequence of effectively

identically shaped narrow voltage spikes, also known as action potentials (AP's). The durations of the interspike intervals (ISI's) between j's AP's constitute a random sequence which we denote by $\{T_k, k = 1, 2, \cdots\}$.

Neuron j's external excitation and inhibition comprises a collection of neural spike trains generated by the afferent presynaptic neurons, called j's afferent cohort. j's afferent synapses are connections between these neurons and j. When a spike arrives at the presynaptic terminal of a member of j's afferent cohort, the depolarization it provides opens the calcium channels in the presynaptic bouton, producing a calcium influx. Calcium elevations in the cytoplasm permit vesicles containing neurotransmitters to fuse quickly with the presynaptic membrane and to release an amount of neurotransmitters based on the size of the vesicle. Those neurotransmitters diffuse within the synaptic cleft. Although some escape, many bind to the postsynaptic receptors. That, in turn, causes the opening of the ion channels, eventually resulting in a net gain of positive or negative voltage across the postsynaptic membrane, either via an excitatory postsynaptic potential (EPSP) or an inhibitory postsynaptic potential (IPSP), respectively. Both EPSP's and IPSP's combine to produce the postsynaptic potential (PSP) at the axon hillock. Once the PSP exceeds a triggering threshold, j generates another output spike that propagates along its axon to each member of a family of neurons referred to as j's efferent cohort. A typical neuron in primate sensory cortex has afferent and efferent cohorts each of which is composed of circa $10,000$ neurons.

As in [6], we allow the weights of the synapses to differ from one another. Now we also admit inhibitory synapses. Furthermore, we model the excitation intensity not presynaptically as in [5] but postsynaptically as the PSP at the axon hillock generated by the arrival thereto and the leakage therefrom of both positively and negatively charged ions. We continue to assume that each of j's refractory periods has the same duration, Δ. This theoretical extension embraces the plasticity of a neuron's synaptic weights, widely considered essential to learning and memory, thereby increasing the chances that the model has practical significance [6] [35].

We model the PSP as a random measure with a continuous-time instantaneous rate function, $\Lambda(t)$, defined by

$$\Lambda(t) := \lim_{\mathrm{d}t \to 0} \frac{\pm Q \cdot \Pr[\text{arrival/departure of ion in}(t, t+\mathrm{d}t)]}{\mathrm{d}t}, \tag{4.3}$$

where Q is the magnitude of the charge of the ion and "+" applies if the ion is either positively charged and is arriving or negatively charged and is departing; otherwise "−" applies. As in [5] [6] [8], we take a time average operation over the rate function $\Lambda(t)$ and obtain

$$\Lambda_k = \frac{1}{T_k - \Delta} \int_{S_{k-1}+\Delta}^{S_k} \Lambda(u)\mathrm{d}u, \tag{4.4}$$

where T_k is the kth ISI duration of j and $S_k = T_1 + \cdots + T_k$.

Henceforth, we suppress the ISI index k and just write T and Λ. Thus, when $\Lambda = \lambda$, the PSP is being assumed to build up according to a diffusion process with a time-averaged intensity λ which must be positive in order for the threshold to be reached.

4.2.3 Information Rate

Here we are interested in the Shannon mutual information, $\mathrm{I}(\Lambda; T)$. Although this has been defined for a single pair of r.v.'s Λ and T, it has been shown that it is a good first-order approximation to the long-term information rate in bits per spike, namely

$$\mathrm{I} := \lim_{N \to \infty} \frac{1}{N} \mathrm{I}(\Lambda_1, \ldots, \Lambda_N; T_1, \ldots, T_N), \tag{4.5}$$

lacking only an information decrement that addresses correlation among successive Λ_i's, which is approximated to the first order in [5] [6] [7]. Since $\tilde{T}$ is a one-to-one function of T, we have $\mathrm{I}(\Lambda; T) = \mathrm{I}(\Lambda; \tilde{T})$, which in turn is defined as

$$\mathrm{I}(\Lambda; \tilde{T}) = \mathbb{E}\left[\log \left(\frac{f_{\tilde{T}|\Lambda}(\tilde{T}|\Lambda)}{f_{\tilde{T}}(\tilde{T})} \right) \right], \tag{4.6}$$

where the expectation is taken with respect to the joint pdf (jpdf) of Λ and $\tilde{T}$. Henceforth, we contract notation from $\tilde{T}$ to T. Toward determining $\mathrm{I}(\Lambda;T)$, we proceed to analyze $f_{T|\Lambda}(t|\lambda)$ and $f_T(t)$ in the GIG case.

4.3 What is being optimized inside the neuron?

Neuron j has the ability to slowly adapt parameters in $f_{T|\Lambda}(t|\lambda)$ by changing synaptic weights via the size and number of presynaptic active zones, the number of vesicles and postsynaptic receptors, among many other properties. In the case of the GIG channel of Eq. (4.1) above, it is able to change the three parameters α, β and γ in the conditional pdf $f_{T|\Lambda}(t|\lambda)$. It is reasonable to postulate that, if the input distribution $f_\Lambda(\lambda)$ changes slowly over time, j will adjust these parameters in an attempt to preserve its bits per joule (bpJ) optimization. From the information theory point of view, it is the joint distribution of Λ and T that determines $\mathrm{I}(\Lambda;T)$. Moreover, $f_{\Lambda,T}(\cdot,\cdot)$ also determines the average energy that j expends, which we model in terms of functions of Λ and T.

Neuron j has a modicum of control over the marginal distribution of Λ, which we measure at the ion level at the axon hillock. For example, j can do this by the aforementioned synaptic plasticity including Long-term potentiation (LTP) and Long-term depression (LTD). However, the bulk of the variations of $f_\Lambda(\cdot)$ is effected by phenomena external to j.

We consider that j's internal modifications to $f_\Lambda(\cdot)$ are in the spirit of a one-to-one transformation of the afferent excitation/inhibition intensity statistics that, in conjunction with a GIG $f_{T|\Lambda}(\cdot|\cdot)$ with appropriately modified values of α, β and γ, facilitates the overall joint bpJ optimization. This formulation differs sharply and significant from our earlier work in which we considered the neuron channel model $f_{T|\Lambda}(\cdot|\cdot)$ to be fixed and solved for which afferent pdf $f_\Lambda(\cdot)$ maximizes bpJ for this fixed channel. That amounted to an assumption that the rest of the brain somehow sees to it that each neuron is stimulated afferently by the $f_\Lambda(\cdot)$ that maximizes its bpJ performance. Now, for each choice of (α,β,γ) we get the

same optimum $f_\Lambda(\cdot)$ we would have the old way (albeit now physically situated at the axon hillock) and hence the same optimized bpJ. The essence of the distinction is that now we are envisioning that there is indeed a "choice" of (α, β, γ) in which j is actively involved. In other words, in response to statistical changes in its external stimulation, j adaptively adjusts both the statistics of the associated $f_\Lambda(\cdot)$ that it produces at the axon hillock and its (α, β, γ) parameters to effect joint bpJ maximization. Mathematically, for each choice of (α, β, γ) there is exactly one $f_\Lambda(\cdot)$ that genuinely maximizes bpJ. Neuron j will not succeed in producing precisely that $f_\Lambda(\cdot)$. However, doing so would not make practical sense because j's $f_\Lambda(\cdot)$ never stays fixed long enough that its precise long-term statistics get the chance to fully exhibit themselves the way a fixed choice of $f_\Lambda(\cdot)$ would for an unadaptable $f_{T|\Lambda}(\cdot|\cdot)$.

4.4 Bits per Joule Optimality Condition

For any choice of $f_\Lambda(\lambda)$, the corresponding pdf of T is given by the integral equation

$$f_T(t) = \int_0^\infty \mathrm{d}\lambda f_{T|\Lambda}(t|\lambda) f_\Lambda(\lambda), \ \forall t \geq 0. \tag{4.7}$$

From Appendix A of [5], we know that $f_{\Lambda,T}(\cdot,\cdot) = f_\Lambda(\cdot) f_{T|\Lambda}(\cdot|\cdot)$ is bpJ-maximizing if and only if the $f_T(\cdot)$ that it generates via Eq. (4.7) satisfies

$$\int_0^\infty \mathrm{d}t f_{T|\Lambda}(t|\lambda) \left[\log\left(\frac{f_{T|\Lambda}(t|\lambda)}{f_T(t)} \right) - \sum_{i=1}^{6} g_i(\lambda, t) \right] = 0, \qquad \forall \lambda \geq 0, \tag{4.8}$$

where $g_i, 1 \leq i \leq 6$, are the energy terms itemized as below:

- $g_1(\lambda, t) = A$, a positive constant. An example of this is the energy j must expend in order to propagate the spike it generates at the end of the ISI to all its target neurons. This and other fixed costs per ISI sum to A.
- $g_2(\lambda, t) = Bt$, where B is a positive constant. The metabolic energy the neuron expends keeping itself healthy and nourished during an ISI would be an example. This grows

linearly with the duration of the ISI no matter whether j is processing relatively intensely during it or not.

- $g_3(\lambda, t) = C\lambda t$, C a positive constant. This term addresses the energy j devotes to generating contributions to the PSP accumulation by processing all the spikes afferent to it during the ISI. [1]

- $g_4(\lambda, t) = \frac{L}{t}$, L being a positive constant. g_4 charges a high energy cost for small values, t, of the r.v. T. That such a small-t penalty must exist is apparent from the definition of Δ as the least amount of time after one AP that another can be produced in response to natural afferent excitation. Although an experimenter can evoke AP's separated by less than Δ by injecting a large current into the axon hillock, this requires supplying power far exceeding what the neuron can expend on its own. It stands to reason that the closer to $t = \Delta$ the neuron produces an AP in response to natural stimulation, the more energy it takes to do so.

- $g_5(\lambda, t) = \frac{G}{\lambda t}$, G being a positive constant. Although $\mathbb{E}[T|\Lambda = \lambda]$ is of the form $Const./\lambda$, so that $\mathbb{E}[\Lambda T] = Const.$, it is still true that no matter how small (large) any λ is, it will occasionally produce a t that also is small (large). g_5 associates an energy expenditure that "corrects" for departures of ΛT from $\mathbb{E}[\Lambda T]$.

- $g_6(\lambda, t) = -D \log t$. If $D > 0$, then this logarithmic term charges an energy penalty whenever $t < 1$ and "rewards" t-values smaller than the time unit, and conversely, if $D < 0$. Note that this logarithmic term is dominated by $g_2(\lambda, t) = Bt$ when t is large and by $g_4(\lambda, t) = L/t$ when t is small, provided D is not many times larger than B or L, respectively. Hence g_6 serves as a fine tuning adjustment to g_2 and/or g_4. It is widely felt that most neurons are finely tuned.

[1] In order for $C\lambda t$ not to be an oversimplification, there needs to be, regardless of the value of λ, enough separate contributions to PSP during an ISI that the coefficient of variation of the number of them and that of the weights of the synapses that generate them both are smaller than, say, 0.1. Primary cortical neurons, among others, conform to these conditions.

Thus, equation (4.8) becomes the following:

$$\int_0^\infty \mathrm{d}t f_{T|\Lambda}(t|\lambda)\left[\log\left(\frac{f_{T|\Lambda}(t|\lambda)}{f_T(t)}\right) - \left(A + Bt + C\lambda t + \frac{G}{\lambda t} + \frac{L}{t}\right) + D\log t\right] = 0, \qquad \forall \lambda \geq 0, \tag{4.9}$$

where the conditional pdf $f_{T|\Lambda}(t|\lambda)$ follows a GIG conditional pdf Eq. (4.1).

4.5 Marginal Output ISI Distribution $f_T(t)$ is a GIG distribution

Temporarily assume that when bpJ is maximized by virtue of satisfaction of Eq. (4.9), the marginal ISI pdf $f_T(t)$ follows a GIG distribution

$$f_T(t) = C_a t^{a-1} \exp\left(-ct - \frac{b}{t}\right), \tag{4.10}$$

where

$$C_a^{-1} = 2\left(\frac{b}{c}\right)^{\frac{a}{2}} K_a(2\sqrt{bc}). \tag{4.11}$$

After substituting Eq. (4.1) and Eq. (4.10) into Eq. (4.9), it follows that

$$\begin{aligned}&\int_0^\infty \mathrm{d}t f_{T|\Lambda}(t|\lambda)\left[\log C_\alpha + \alpha\log\lambda - A + (\alpha - 1 + D)\log t - Bt - (\gamma + C)\lambda t - \frac{(\beta + G)/\lambda + L}{t}\right] \\ = &\int_0^\infty \mathrm{d}t f_{T|\Lambda}(t|\lambda)\left[\log C_a + (a-1)\log t - ct - \frac{b}{t}\right], \ \forall \lambda \geq 0.\end{aligned} \tag{4.12}$$

Next, note the following expectation equalities:

$$\begin{aligned}\mathbb{E}[T|\lambda] &= \int_0^\infty \mathrm{d}t\ t f_{T|\Lambda}(t|\lambda) \\ &= \int_0^\infty \mathrm{d}t\ C_\alpha \lambda^\alpha t^\alpha \exp\left(-\gamma\lambda t - \frac{\beta}{\lambda t}\right) \\ &= \sqrt{\frac{\beta}{\gamma}}\frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\frac{1}{\lambda};\end{aligned} \tag{4.13}$$

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\middle|\lambda\right] &= \int_0^\infty \mathrm{d}t\ \frac{1}{t} f_{T|\Lambda}(t|\lambda) \\ &= \int_0^\infty \mathrm{d}t\ C_\alpha \lambda^\alpha t^{\alpha-2} \exp\left(-\gamma\lambda t - \frac{\beta}{\lambda t}\right) \\ &= \sqrt{\frac{\gamma}{\beta}}\frac{K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\lambda;\end{aligned} \tag{4.14}$$

$$\begin{aligned}\mathbb{E}[\log T|\lambda] =& \mathbb{E}[\log \lambda T|\lambda] - \log\lambda \\ \overset{a}{=}& \int_0^\infty \mathrm{d}x\ C_\alpha x^{\alpha-1}\exp\left(-\gamma x - \frac{\beta}{x}\right)\log x - \log\lambda \\ =& \frac{1}{2}\log\frac{\beta}{\gamma} + \frac{\frac{\partial}{\partial\alpha}[K_\alpha(2\sqrt{\beta\gamma})]}{K_\alpha(2\sqrt{\beta\gamma})} - \log\lambda.\end{aligned} \tag{4.15}$$

Equality (a) is due to the replacement of λt with x.

Rewriting Eq. (4.12) as

$$\begin{aligned}\log C_\alpha + \alpha\log\lambda - \log C_a - A + (\alpha+D-a)\mathbb{E}[\log T|\lambda] - (B-c+(\gamma+C)\lambda)\mathbb{E}[T|\lambda] \\ - \left(L - b + \frac{\beta+G}{\lambda}\right)\mathbb{E}\left[\frac{1}{T}\middle|\lambda\right] = 0, \forall\lambda \geq 0,\end{aligned} \tag{4.16}$$

and substituting Eqs. (4.13), (4.14) and (4.15) into Eq. (4.16) yields

$$\log C_\alpha + \alpha \log \lambda - \log C_a - A + (\alpha + D - a)\left(\frac{1}{2}\log\frac{\beta}{\gamma} + \frac{\frac{\partial}{\partial\alpha}[K_\alpha(2\sqrt{\beta\gamma})]}{K_\alpha(2\sqrt{\beta\gamma})} - \log\lambda\right)$$
$$- (B - c + (\gamma + C)\lambda) \times \left(\sqrt{\frac{\beta}{\gamma}}\frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\frac{1}{\lambda}\right) - \left(L - b + \frac{\beta + G}{\lambda}\right)\left(\sqrt{\frac{\gamma}{\beta}}\frac{K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\lambda\right) = 0, \tag{4.17}$$

Upon defining

$$\begin{aligned} C_{const} :=& \log C_\alpha - \log C_a - A \\ =& \frac{a}{2}\log\frac{b}{c} - \frac{\alpha}{2}\log\frac{\beta}{\gamma} + \log\frac{K_a(2\sqrt{bc})}{K_\alpha(2\sqrt{\beta\gamma})} - A. \end{aligned} \tag{4.18}$$

We may recast Eq. (4.17) in the form

$$\begin{aligned} C_{const} + \alpha\log\lambda + (\alpha + D - a)\left(\frac{1}{2}\log\frac{\beta}{\gamma} + \frac{\frac{\partial}{\partial\alpha}[K_\alpha(2\sqrt{\beta\gamma})]}{K_\alpha(2\sqrt{\beta\gamma})} - \log\lambda\right) - (B - c + (\gamma + C)\lambda) \\ \times \left(\sqrt{\frac{\beta}{\gamma}}\frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\frac{1}{\lambda}\right) - \left(L - b + \frac{\beta + G}{\lambda}\right)\left(\sqrt{\frac{\gamma}{\beta}}\frac{K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\lambda\right) = 0, \ \forall\lambda \geq 0. \end{aligned} \tag{4.19}$$

By letting $a = D$, $b = L$ and $c = B$, all λ's disappear entirely from Eq. (4.19), yielding

$$C'_{const} + \alpha\frac{\frac{\partial}{\partial\alpha}K_\alpha(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})} - (\gamma + C)\sqrt{\frac{\beta}{\gamma}}\frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})} - (\beta + G)\sqrt{\frac{\gamma}{\beta}}\frac{K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})} = 0, \ \forall\lambda \geq 0, \tag{4.20}$$

where

$$C'_{const} = \frac{a}{2}\log\frac{b}{c} + \log\frac{K_a(2\sqrt{bc})}{K_\alpha(2\sqrt{\beta\gamma})} - A.$$

Therefore, our temporary assumption that the bpJ-maximizing marginal output ISI distribution is GIG is validated because the bpJ-maximizing $f_T(t)$ is the GIG pdf

$$f_T(t) = C_a t^{a-1} \exp\left(- ct - \frac{b}{t} \right) \tag{4.21}$$

with C_a defined as in Eq. (4.11) and with $a = D$, $b = L$ and $c = B$.

4.6 Shannon Mutual Information $I(\Lambda; T)$ Calculation

In order to calculate the Shannon mutual information $I(\Lambda; T)$, we first need a lemma.

Lemma 1 *Channel 1 and Channel 2 described below are equivalent in the sense of probabilistically generating the same output r.v. T given the identical input distribution $f_\Lambda(\lambda)$.*

Channel 1: The r.v.'s Λ and T are related by

$$T \stackrel{d}{=} \frac{1}{\Lambda} \cdot U, \tag{4.22}$$

where U is independent of Λ and has the pdf

$$f_U(u) = C_\alpha u^{\alpha-1} \exp\left(- \gamma u - \frac{\beta}{u} \right), 0 < u < \infty, \tag{4.23}$$

which is the special case of the conditional pdf $f_{T|\Lambda}(t|\lambda)$ in which $\lambda = 1$ and u plays the role of t. Eq. (4.22) is equality in distribution, which implies that T and $\frac{U}{\Lambda}$ are identically distributed r.v.'s.

Channel 2: The property characterizing this channel is the conditional pdf

$$f_{T|\Lambda}(t|\lambda) = C_\alpha \lambda^\alpha t^{\alpha-1} \exp\left(- \gamma \lambda t - \frac{\beta}{\lambda t} \right), \tag{4.24}$$

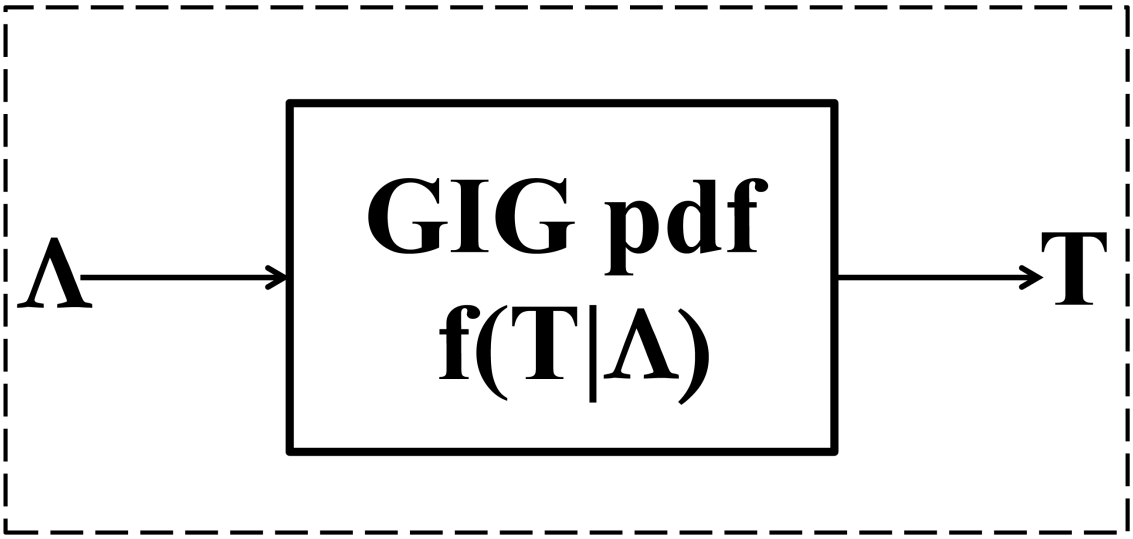

Figure 4.3: GIG Channel 1.

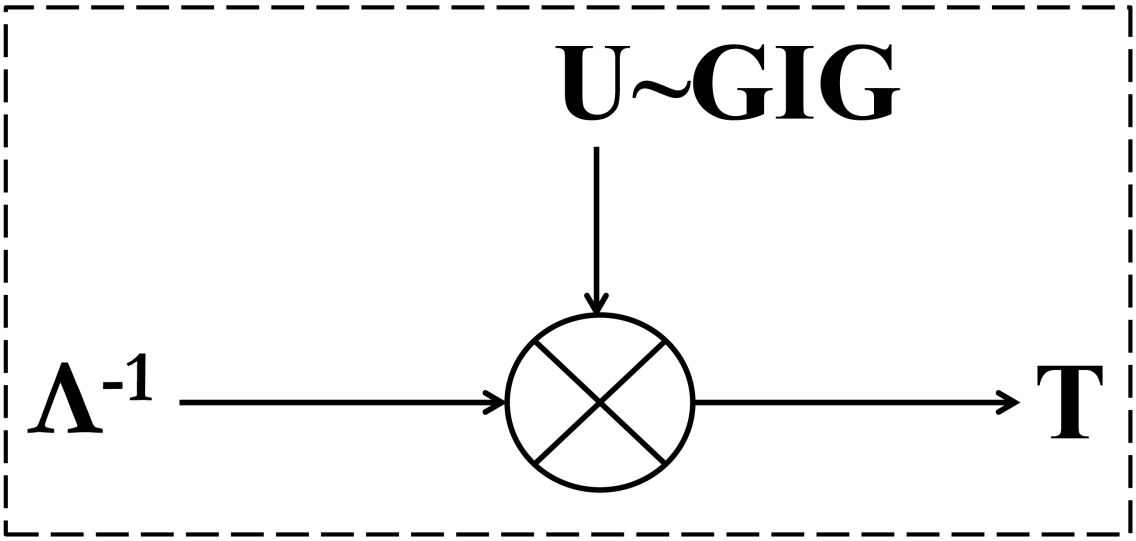

Figure 4.4: GIG Channel 2.

where C_α is defined as in Eq. (4.2).

Proof: Channel 1 $\Rightarrow$ Channel 2.

For Channel 1, suppose $\Lambda = \lambda$, then $T = \frac{U}{\lambda}$. Hence we have the following

$$|f_{T|\Lambda}(t|\lambda)\mathrm{d}t| = \left| f_{\frac{U}{\lambda}|\Lambda}\left(\frac{u}{\lambda}\middle|\lambda\right)\mathrm{d}t\right|, \tag{4.25}$$

$$f_{T|\Lambda}(t|\lambda) \overset{a}{=} f_{\frac{U}{\lambda}}\left(\frac{u}{\lambda}\right). \tag{4.26}$$

Equality (a) holds due to the independency between U and Λ and the nonnegativity of the pdf's.

Since r.v. U follows a GIG distribution with pdf Eq. (4.23), it yields that

$$\left|f_{\frac{U}{\lambda}}(x)\mathrm{d}x\right| = |f_U(u)\mathrm{d}u| = |\lambda f_U(\lambda x)\mathrm{d}x|, \tag{4.27}$$

$$f_{\frac{U}{\lambda}}(x) \overset{b}{=} \lambda f_U(\lambda x). \tag{4.28}$$

Equality (b) holds because of the nonnegativity of the pdf's.

Thus, based on Eqs. (4.26), (4.28) and (4.23), we finally have

$$f_{T|\Lambda}(t|\lambda) = \lambda \cdot C_\alpha(\lambda t)^{\alpha-1} \exp\left(-\gamma\lambda t - \frac{\beta}{\lambda t}\right). \tag{4.29}$$

Channel 2 $\Rightarrow$ Channel 1.

For Channel 2, we have as the conditional pdf describing the nature of the channel noise

$$f_{T|\Lambda}(t|\lambda) = \lambda \cdot C_\alpha(\lambda t)^{\alpha-1} \exp\left(-\gamma\lambda t - \frac{\beta}{\lambda t}\right). \tag{4.30}$$

Let $U = \lambda \cdot T$ (i.e., $u = \lambda t$); then

$$\begin{aligned} |f_{T|\Lambda}(t|\lambda)\mathrm{d}t| &= |f_{U|\Lambda}(u|\lambda)\mathrm{d}u| \\ &= |f_{U|\Lambda}(u|\lambda)\lambda\mathrm{d}t|. \end{aligned} \tag{4.31}$$

Furthermore, if we assume U is independent of Λ, we have

$$f_U(u) = f_{U|\Lambda}(u|\lambda) = \frac{1}{\lambda} \cdot f_{T|\Lambda}(t|\lambda) \tag{4.32}$$

$$= C_\alpha u^{\alpha-1} \exp\left(-\gamma u - \frac{\beta}{u} \right). \tag{4.33}$$

Therefore, Channel 1 and Channel 2 are equivalent because they share the same condional pdf of T given Λ. $\square$

According to Lemma 1, by defining $W = \log U$, $V = \log T$ and $Z = -\log \Lambda$, we obtain

$$V = W + Z, \tag{4.34}$$

where W and Z are independent.

Henceforth, it follows

$$\begin{aligned} I(\Lambda; T) &= I(Z; V) \\ &= h(V) - h(V|Z) \\ &= h(\log T) - h(\log U). \end{aligned} \tag{4.35}$$

Since

$$\begin{aligned} h(\log T) &= -\mathbb{E}[\log f_V(V)] = -\mathbb{E}[\log (f_T(T) \cdot T)] \\ &= -\mathbb{E}[\log f_T(T)] - \mathbb{E}[\log T] \\ &= h(T) - \mathbb{E}[\log T], \end{aligned} \tag{4.36}$$

the mutual information $I(\Lambda;T)$ then can be written as

$$\begin{aligned} I(\Lambda;T) &= [h(T) - \mathbb{E}\log T] - [h(U) - \mathbb{E}\log U] \\ &= -\int_0^\infty \mathrm{d}t f_T(t)\log\left(t \cdot f_T(t)\right) + \int_0^\infty \mathrm{d}u f_U(u)\log\left(u \cdot f_U(u)\right). \end{aligned} \tag{4.37}$$

Due to the following integral equalities:

$$\int_0^\infty \mathrm{d}u\ C_\alpha u^{\alpha-1}\exp\left(-\gamma u - \frac{\beta}{u}\right)\cdot \alpha\log u = \alpha\left[\frac{\frac{\partial}{\partial\alpha}K_\alpha(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})} + \frac{1}{2}\ln\frac{\beta}{\gamma}\right]; \tag{4.38}$$

$$\int_0^\infty \mathrm{d}u\ C_\alpha u^{\alpha-1}\exp\left(-\gamma u - \frac{\beta}{u}\right)\cdot \gamma u = \gamma\sqrt{\frac{\beta}{\gamma}}\frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}; \tag{4.39}$$

$$\int_0^\infty \mathrm{d}u\ C_\alpha u^{\alpha-1}\exp\left(-\gamma u - \frac{\beta}{u}\right)\cdot \frac{\beta}{u} = \beta\sqrt{\frac{\gamma}{\beta}}\frac{K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}, \tag{4.40}$$

by substituting Eqs. (4.38) (4.39) (4.40) into Eq. (4.37) we can eventually obtain

$$\begin{aligned} I(\Lambda;T) =& \log\frac{K_a(2\sqrt{bc})}{K_\alpha(2\sqrt{\beta\gamma})} + \sqrt{bc}\left[\frac{K_{a+1}(2\sqrt{bc}) + K_{a-1}(2\sqrt{bc})}{K_a(2\sqrt{bc})}\right] - \sqrt{\beta\gamma}\left[\frac{K_{\alpha+1}(2\sqrt{\beta\gamma}) + K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\right] \\ &+ \alpha\frac{\frac{\partial}{\partial\alpha}K_\alpha(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})} - a\frac{\frac{\partial}{\partial a}K_a(2\sqrt{bc})}{K_a(2\sqrt{bc})}. \end{aligned} \tag{4.41}$$

4.7 Energy Expenditure $e(\Lambda, T)$ Calculation

For the expected energy expenditure, we have

$$e(\Lambda, T) = \mathbb{E}\left[\sum_{i=1}^{6} G_i(\Lambda, T)\right] \tag{4.42}$$

$$= A + B\mathbb{E}[T] + L\mathbb{E}\left[\frac{1}{T}\right] + C\mathbb{E}[\Lambda T] + G\mathbb{E}\left[\frac{1}{\Lambda T}\right] - D\mathbb{E}[\log T]. \tag{4.43}$$

Since

$$\mathbb{E}[T] = \sqrt{\frac{b}{c}}\frac{K_{a+1}(2\sqrt{bc})}{K_a(2\sqrt{bc})}; \tag{4.44}$$

$$\mathbb{E}\left[\frac{1}{T}\right] = \sqrt{\frac{c}{b}}\frac{K_{a-1}(2\sqrt{bc})}{K_a(2\sqrt{bc})}; \tag{4.45}$$

$$\mathbb{E}[\log T] = \frac{\frac{\partial}{\partial a}K_a(2\sqrt{bc})}{K_a(2\sqrt{bc})} + \frac{1}{2}\ln\frac{b}{c}; \tag{4.46}$$

$$\mathbb{E}[\Lambda T] = \sqrt{\frac{\beta}{\gamma}}\frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}; \tag{4.47}$$

$$\mathbb{E}\left[\frac{1}{\Lambda T}\right] = \sqrt{\frac{\gamma}{\beta}}\frac{K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}, \tag{4.48}$$

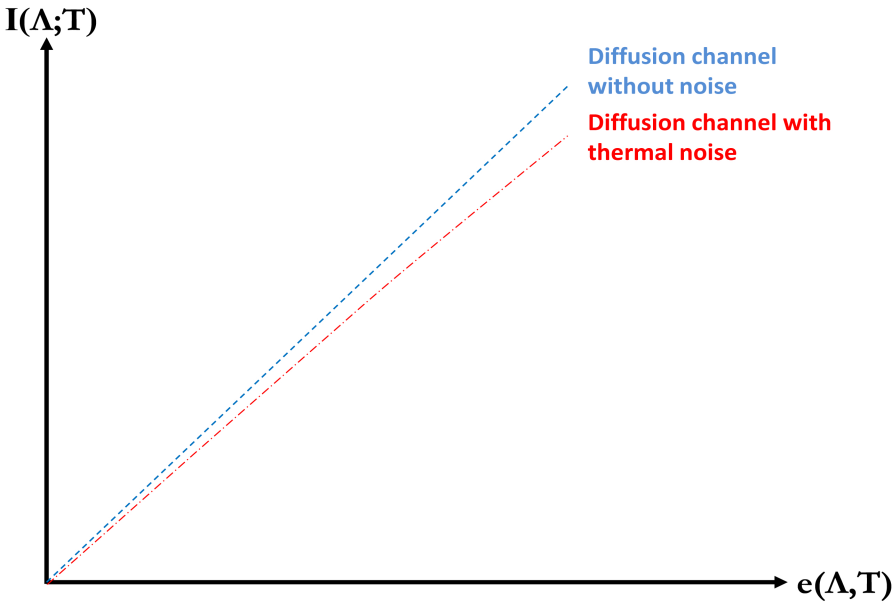

Figure 4.5: Information vs energy plot.

the expected energy consumption can be written as

$$e(\Lambda,T) = A + \sqrt{bc}\left[\frac{K_{a+1}(2\sqrt{bc}) + K_{a-1}(2\sqrt{bc})}{K_a(2\sqrt{bc})}\right] + C\sqrt{\frac{\beta}{\gamma}}\frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})} + G\sqrt{\frac{\gamma}{\beta}}\frac{K_{\alpha-1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})} - a\left[\frac{\frac{\partial}{\partial a}K_a(2\sqrt{bc})}{K_a(2\sqrt{bc})} + \frac{1}{2}\ln\frac{b}{c}\right]. \tag{4.49}$$

Substituting Eq. (4.20) into Eq. (4.49), we can surprisingly find

$$I(\Lambda;T) = e(\Lambda,T). \tag{4.50}$$

4.8 Possible Explanations for $I(\Lambda;T) = e(\Lambda,T)$

Equation (4.50) establishes the perhaps surprising result that, under the assumptions employed to derive it, the curve of the optimum number of bits per second transmitted versus the average power the GIG neuron expends in order to send them actually doesn't curve at all; it

is a straight line! (It appears to have slope 1, but its numerical slope actually depends on the units used for bits and for energy.) Although such a straight line does not violate information theory's dictate that an average power constrained channel's capacity in bits per second must be concave, it certainly is unintuitive that said curve actually would be a straight line in practice. In the following paragraphs, we give some reasons for why the capacity curve for a GIG neuron channel actually is strictly concave.

Our GIG diffusion model assumes the existence of a probability density solution to the optimization problem of maximizing bits per Joule. If this assumption is invalid, then such a Lagrangian formulation of the problem is inappropriate. If the optimum occurs on the constraint boundary rather than in the interior, it becomes necessary to use Karush-Kuhn-Tucker (KKT) theory to solve the problem via convex mathematical programming. In that event the information, I, versus energy, e, curve likely will bend eventually as bits and energy increase instead of remaining a straight line. Another reason why the physical I versus e curves likely will bend eventually is that the GIG diffusion model we have analyzed does not take thermal noise into account, whereas real neuron channels always harbor it. Furthermore, the Berger-Levy (BL) model [5] does not have a $\frac{1}{T}$ term in the exponent the way the GIG model does. Accordingly, the BL model scarcely discourages high values of the frequency variable λ, whereas the GIG diffusion channel imposes a rapid decay of $f_\Lambda(\lambda)$ for large λ's. Real neurons can and do open up their bandwidth in order to process rapidly varying signals (i.e., high-λ signals). Because of this adaptive bandwidth variation, thermal noise power goes up whenever the neuron is obliged to process high frequency excitations because thermal noise fills the entire broad bandwidth associated with high frequency signal processing. Accordingly, for white thermal noise of power spectral density N_0, the signal-to-noise ratio, SNR=$\frac{S}{N_0 \cdot \lambda}$, decreases as λ increases because the average energy, S, does not increase when λ increases. Not only is the energy cost of propagating a spike to all of the neuron's targets the same whether the ISI at whose end the spike is generated is a short ISI typical of a high λ or a long one characteristic of a low λ, but also the average amount of energy expended to process

afferent signals until the PSP crosses the threshold does not depend on λ either for a GIG neuron. This is because that energy is proportional to the λt product, so by the smoothing property of conditional expectation its expected value is $\mathbb{E}[\Lambda T] = \mathbb{E}[\Lambda \mathbb{E}[T|\Lambda]]$; however, for a GIG model $\mathbb{E}[T|\Lambda]$ is proportional to $1/\Lambda$, so the overall value of $\mathbb{E}[\Lambda T]$ is a constant. Accordingly, we conjecture that in order for a neural network to be able to perform close to the thermodynamic limit, it needs to excite its neurons so that each of them operates principally in the (low bits, low energy) regime most of the time and is driven into the (high bits, high energy) regime relatively rarely, thus allowing the brain to operate as a totality in an energy-efficient manner.

However, the relatively rare but often important instances during which a neuron's input rate λ is well above its mean value are characterized by the following two scenarios. First, when low SNR occurs, the percentage accuracy of the bits received during such instances (not the number of bits received, but their percentage accuracy) will decrease due to the fact that such low-SNR events suffer from reduced bits transmitted successfully per unit of energy expended, thereby probably resulting in an energy-constrained capacity curve that exhibits strict concavity. Alternatively, if one wants to maintain the accuracy to recover the received bits, it needs extra energy to amplify the signal in order to increase the SNR, which will eventually cause the curve to bend.

4.9 Bits per Joule Optimizing Input Probability Distribution $f_\Lambda(\lambda)$

According to Lemma 1, by defining $W = \log U$, $V = \log T$ and $Z = -\log \Lambda$, we obtain

$$V = W + Z. \tag{4.51}$$

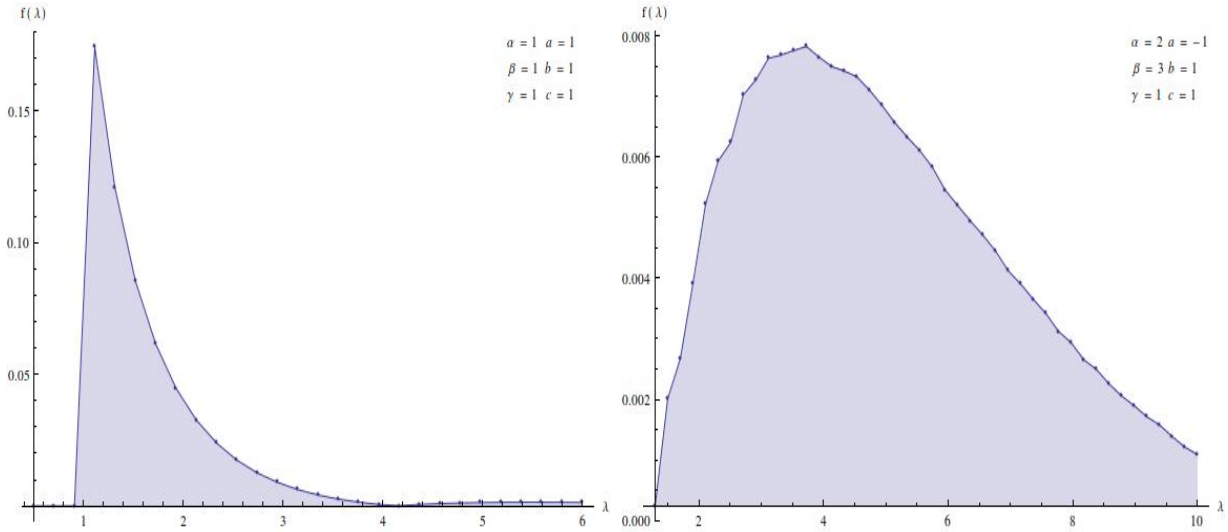

Figure 4.6: $f_\Lambda(\lambda)$ plots for two sets of $(\alpha, \beta, \gamma, a, b, c)$.

According to Equation 7 on page 959 of [36],

$$K_\nu(xz) = \frac{z^2}{2}\int_0^\infty dt \exp\left[-\frac{1}{2}\left(xt + \frac{xz^2}{t}\right)\right] t^{-\nu-1}, \tag{4.52}$$

the characteristic functions of V and W can be obtained as

$$\psi_V(x) = \left(\frac{b}{c}\right)^{\frac{jx}{2}} \frac{K_{-a-jx}(2\sqrt{bc})}{K_a(2\sqrt{bc})}; \tag{4.53}$$

$$\psi_W(x) = \left(\frac{\beta}{\gamma}\right)^{\frac{jx}{2}} \frac{K_{-\alpha-jx}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}. \tag{4.54}$$

Therefore, due to the independence between V and W, the characteristic function of Z can be written as:

$$\begin{aligned}\psi_Z(x) &= \frac{\psi_V(x)}{\psi_W(x)} \\ &= \left(\frac{\gamma b}{\beta c}\right)^{\frac{jx}{2}} \frac{K_\alpha(2\sqrt{\beta\gamma})}{K_a(2\sqrt{bc})} \frac{K_{-a-jx}(2\sqrt{bc})}{K_{-\alpha-jx}(2\sqrt{\beta\gamma})},\end{aligned} \tag{4.55}$$

which can be Fourier inverted allowing us to eventually obtain $f_\Lambda(\lambda)$ as

$$f_\Lambda(\lambda) = \frac{1}{2\pi\lambda}\int_{-\infty}^{\infty} \mathrm{d}x \left(\lambda\sqrt{\frac{\gamma b}{\beta c}}\right)^{jx} \frac{K_{-a-jx}(2\sqrt{bc})}{K_{-\alpha-jx}(2\sqrt{\beta\gamma})} \tag{4.56}$$

Through numerical calculations, two instances of $f_\Lambda(\lambda)$ are illustrated in Figure 4.6. For certain combinations of $(\alpha, \beta, \gamma, a, b, c)$, the integral in Eq. (4.56) does not converge. Then KKT analysis (see Section 4.8) must be employed and may lead to probability density functions (pdf's) having to be displaced by cumulative density functions (cdf's) that possess atoms in the high-λ region.

4.10 Summary

We have shown that, when neuron j employs a GIG conditional distribution as the channel density function and bpJ-maximizing is achieved, the output ISI distribution is a related GIG marginal distribution. This has allowed us to compute the tradeoff between information rate and average power in the GIG class in a markedly simpler way requiring only a one dimensional integral instead of several multidimensional integrals. Also, we show how to numerically obtain the associated input distribution $f_\Lambda(\lambda)$. By generalizing from the Gamma and IG families to the GIG family, the derived results contain [5] [6] [7] [8] as special cases in which the three parameters are set to specific values.

Chapter 5

Experimental and Biophysical Neuron Modeling: The Paradoxical Energy Efficiency of Retinothalamic Transmission

All science is either physics or stamp collecting.

-Ernest Rutherford

5.1 Problem Statement

In the early visual processing pathways of mammalian brains, action potential (AP) generated by a retinal ganglion cell (RGC) must first activate thalamic relay cells in the lateral geniculate nucleus (LGN) in order for the RGC to influence the primary visual cortex (V1). However, during active vision, only a fraction of the spikes traveling along the optic nerve can successfully activate a given downstream LGN cell, leading to a 2- to 4-fold reduction in the total number of spikes across the retinothalamic synaptic connection [55]. According to Barlow's principles of efficient coding [52], a minimum number of impulses should be

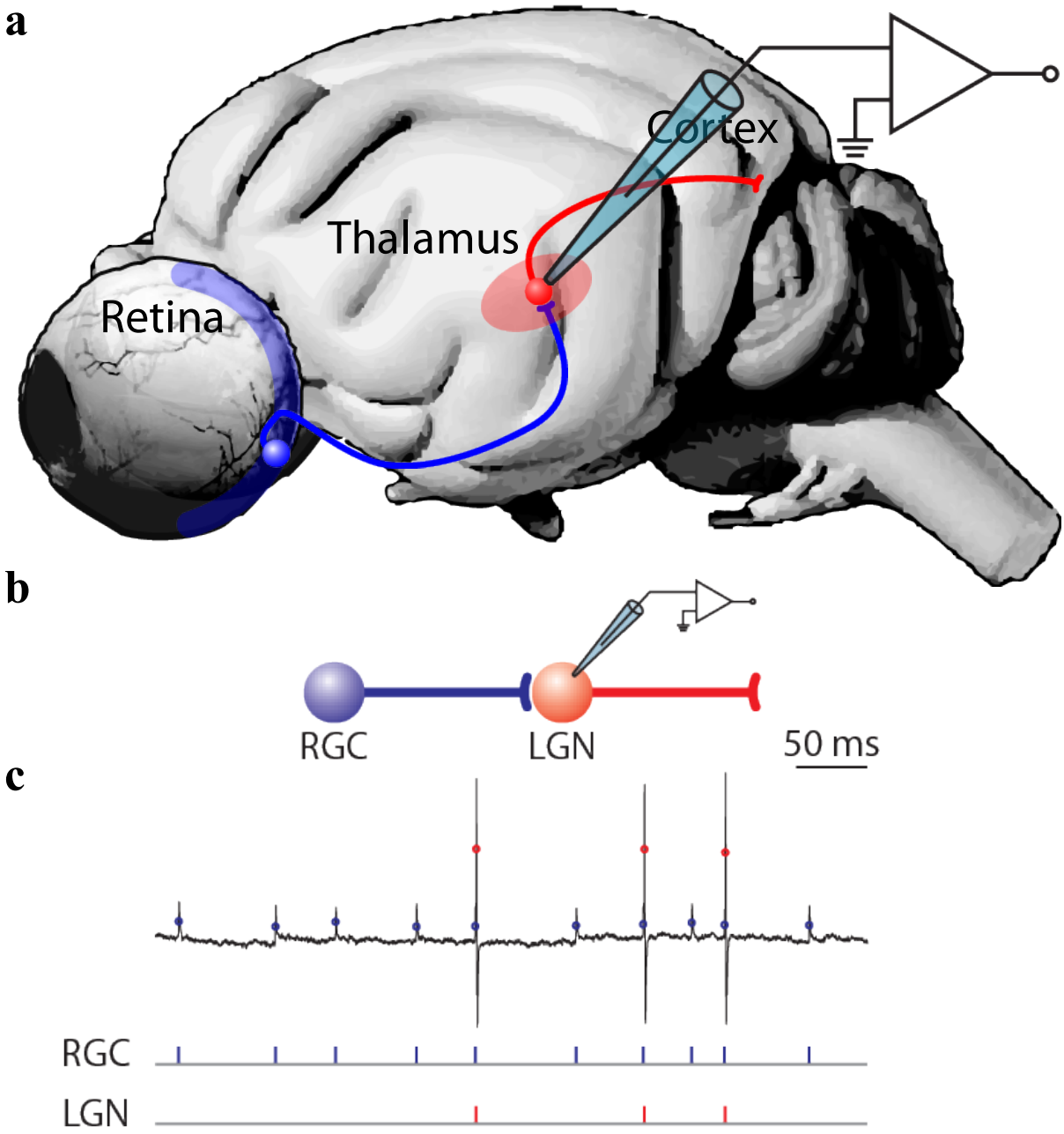

Figure 5.1: Retinothalamic transmission.

used to encode an information source, which implies that the retinal and thalamic neural codes cannot be equally efficient. The retinothalamic synapse selectively relays only the most informative retinal spikes, thereby preserving the most important information [53] [54] [55].

It is unclear why RGCs generate redundant and apparently unnecessary spikes, as this entails an inefficient use of energy in the brain. Here, we resolve this apparent paradox by analyzing simultaneous RGC and LGN recordings [55] with the Berger-Levy energy efficiency theory of neural computation and communication [5] [6]. Our analysis suggests that the retinal spike code, despite its relative inefficiency compared to the thalamic spike code, is an energy-efficient substrate for generating the thalamic code since it maximizes the information transmitted to the cortex per unit of expended energy.

Signaling by the all-or-none spikes used by neurons to transmit information over a long distance and the subsequent release of neurotransmitter at synapses is expensive [23] [27] [4]. In every second 10^{15} synaptic signals are transmitted in the human brain and a single message at a chemical synapse consumes 10^5 ATP molecules [20]. Although some neurons minimize the energy they need to function [4], there appears to be an exception at an early stage of visual processing.

In the mammalian visual pathway, spike trains from the retina traverse the optic nerves and are then relayed into the lateral geniculate nucleus (LGN), from which information subsequently projects to the primary visual cortex (V1). The LGN is called a relay station because the spikes coming from the retinal ganglion cells (RGCs) obligatorily activate the thalamic relay cells in order to activate downstream V1 neurons (Fig. 5.1a). Each LGN cell receives inputs from only a few RGCs with one that is dominant, a near-one-to-one mapping (Fig. 5.1b), and there are 2 to 4 times more retinal input spikes than output LGN spikes (Fig. 5.1c). This leads to a paradox: If the retinal code is optimally efficient, then the LGN would be losing a significant amount of information by relaying only a small fraction of retinal spikes; on the other hand, if the thalamic code is efficient, then the retina would be wastefully generating many redundant spikes that convey little information.

5.2 Berger-Levy Energy Efficient Theory of Neural Computation and Communication

In communicating information, it is important to consider the energy expenditure as well as the information transmitted. That is, in order to transmit information in an energy efficient manner, it is necessary to maximize bits communicated per joule of energy expended. From this viewpoint the objective function to be optimized is information (I) transmitted per energy (e) expended:

$$\text{Maximize} \quad I_j = \frac{I}{e} \frac{bits}{joule}. \tag{5.1}$$

To study the retinothalamic synaptic transmission quantitatively, we used the Berger-Levy (B-L) theory of neural computation and communication, which was originally based on an integrate-and-fire model [5] and subsequently generalized to variable EPSP amplitudes and decay [6] (Fig. 5.2). The B-L theory predicts optimal input and output probability distributions that maximize the bits of information a neuron conveys to its efferent targets per joule of energy it expends for action potential (AP) generation, postsynaptic accumulation and basal metabolism. The theory implies (1) that the intervals between consecutive output spikes must be distributed according to a gamma distribution with shape and scale parameters denoted by κ and b, respectively, and (2) the scaled reciprocal of the input intensity must be distributed according to a beta distribution with three free parameters, two of them being the aforementioned κ and b and the third being a threshold-related parameter, m. In order to quantitatively compare the B-L theory with experimental observations, the spike train statistics from simultaneous recordings of synaptically-connected RGCs and LGN cells [55] were analyzed here using statistical fitting methods (see Section 5.4).

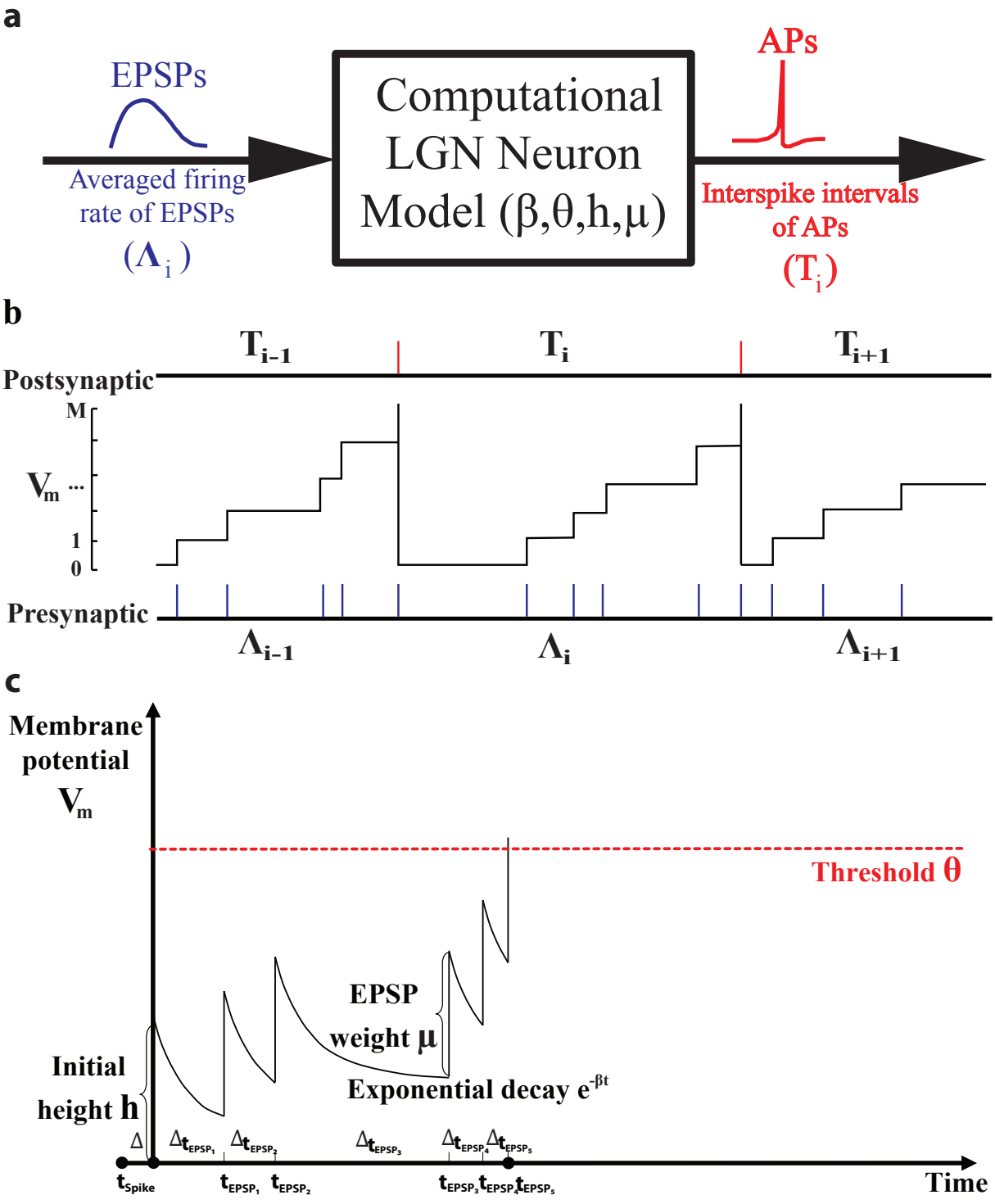

Figure 5.2: LGN neuron model.

5.3 Statistical Analysis

Figure 5.1 illustrates the experimental setup with electrophysiological recordings simultaneously measuring the spikes coming from the retina and the LGN while stimulating ganglion cells with retinal inputs [55]. Figure 3a provides an example of a single thalamic relay cell that has been reasonably well fitted with parameters $\kappa = 0.6703$ and $b = 13.4601$ using maximum-likelihood (ML) estimation. After performing a ML data fitting for a few dozen thalamic relay cells, we consistently observed a cloud of values for the parameters, κ and b, with median centered on $\kappa = 0.72$ and $b = 7.00$ (Fig. 3b). The concentration of κ between 0.6 and 0.8 suggests non-Poisson statistics, making small interspike interval (ISI) more probable compared to a Poisson case with $\kappa = 1$, excluding the refractory period.

Fitting spiking data of a typical retinal cell with statistical ML estimation of the threshold parameter (m_{ML}) is illustrated for a particular ganglion cell (Fig. 5.3c). We then compared the values of m_{ML} with the experimental spiking ratio of RGC over LGN (m_E) across the recorded retinal cell population (Fig. 5.3d), and observed that the two estimates are well correlated but differ systematically, possibly due to the fact that a simple integrate-and-fire neuron model was utilized. The median of the population distribution (Fig. 5.3d, red cross) indicates that the estimated number of inputs within one ISI lies between 2 and 4. This quantitatively reveals how the thalamic relay cell is processing incoming retinal spikes on average. When 2 to 4 spikes occur close to each other in time, the threshold is reached and the information is relayed to the cortex. This suggests that the LGN is not merely a passive relay but rather that it exhibits characteristics of a temporal filter for selectively transmitting information from RGC to V1.

Our model specifically predicts conditions on energy costs in neurons for optimal I_j when the number of retinal spikes is 2- to 4-fold greater than the number of thalamic spikes, as observed experimentally. The expected unit energy cost, e, within a single output ISI has

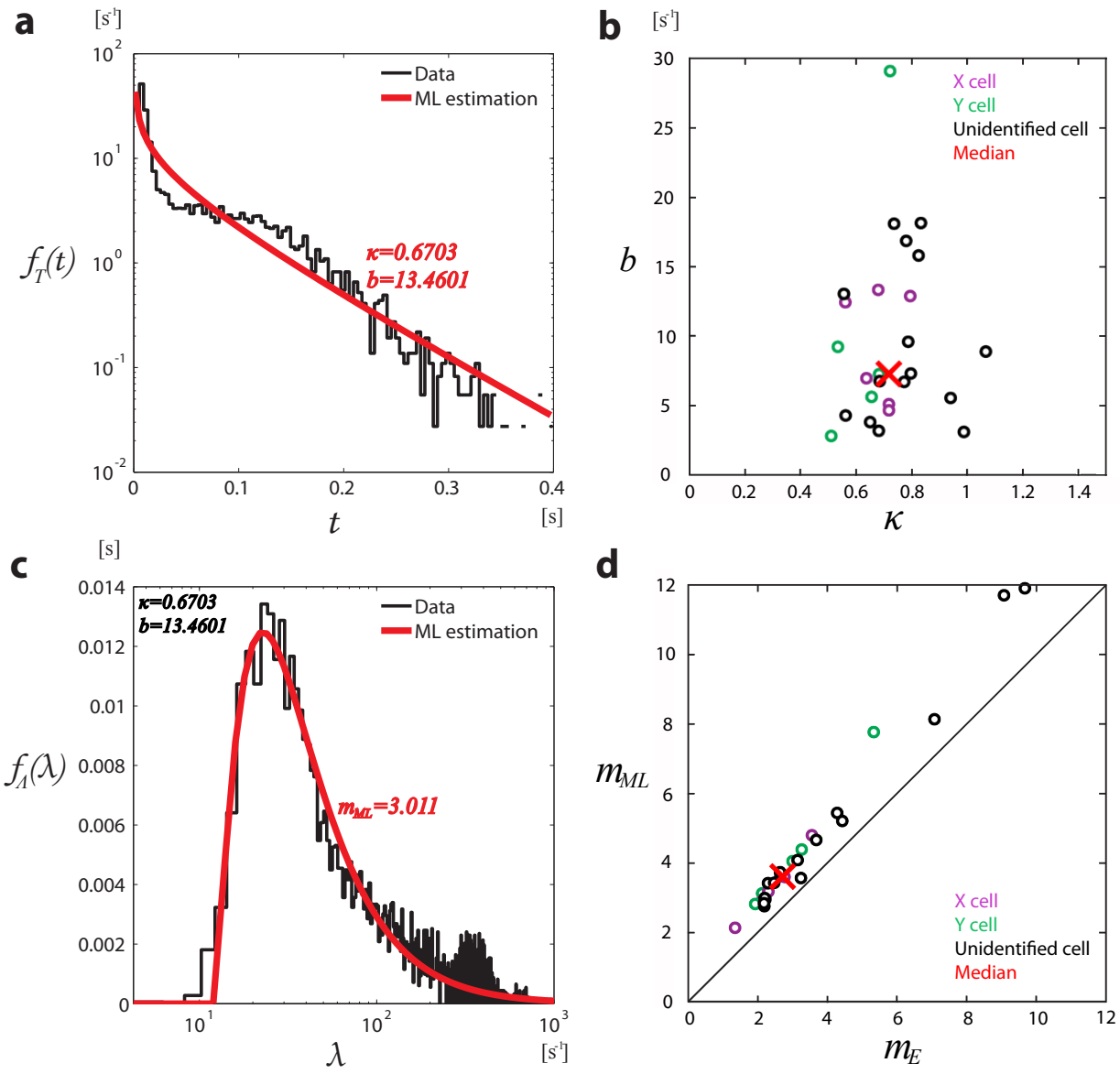

Figure 5.3: Thalamic interspike interval (ISI) parameter estimation and retinal averaged firing rate parameter estimation.

three parts:

$$e = 1 + \rho m + \sigma, \tag{5.2}$$

where 1 is the normalized energy cost of one LGN spike; m is the average number of retinal inputs needed for the LGN cell to cross the threshold; ρ represents the energy cost ratio

between the processing of a single input at an LGN synapse and the producing of an LGN spike; and ρ represents the energy cost ratio between the metabolism during one LGN ISI and the producing of a LGN spike. Although ρ and σ are not known empirically, BL theory predicts combinations that achieve the optimal ratio m with respect to the maximized bits per joule, I_j (see Fig. S4). A particular case in which between 2 and 4 RGC spikes within a short time interval are required to produce an LGN spike is illustrated in Figure 5.4 when $\rho = 1$ and $\sigma = 0$. For this case, I_j is optimal when the amount of energy expended in an LGN spike is roughly equal to the amount of energy in processing an input at an LGN synapse ($\rho = 1$). In another possible optimal case ($\rho = 2$ and $\sigma = 2$) (See Fig. S4), the amount of energy in processing an input at an LGN synapse is the same as the metabolic energy during one ISI and both are twice the cost of an LGN spike. These predictions could be tested by measuring the energy cost for each step in transferring information from a RGC to an LGN cell.

Feedback inhibition from V1 could further contribute to the drop in LGN spiking rate through inhibition from the reticular nucleus of the thalamus, thereby reducing temporal redundancy and saving energy. It would be helpful to measure both feedforward and feedback inhibitory signaling in the LGN [59] [60]. Cortical regions might have their own time-varying viewpoint regarding which information entering the LGN might be most needed and which information can be ignored from moment to moment. The B-L theory should apply to these cases as well. However we must proceed cautiously in attempting to generalize these RGC-LGN findings to other cortical areas, since the tens of thousands of inputs afferent to cortical neurons are correlated and intertwined with one another.

Optimizing bits per joule rather than pure information transfer, which accounts for the 2- to 4-fold greater number of retinal spikes than of thalamic spikes, resolves the apparent paradox of retinothalamic synaptic transmission and could be a general principle for studying information transmission elsewhere in the brain.

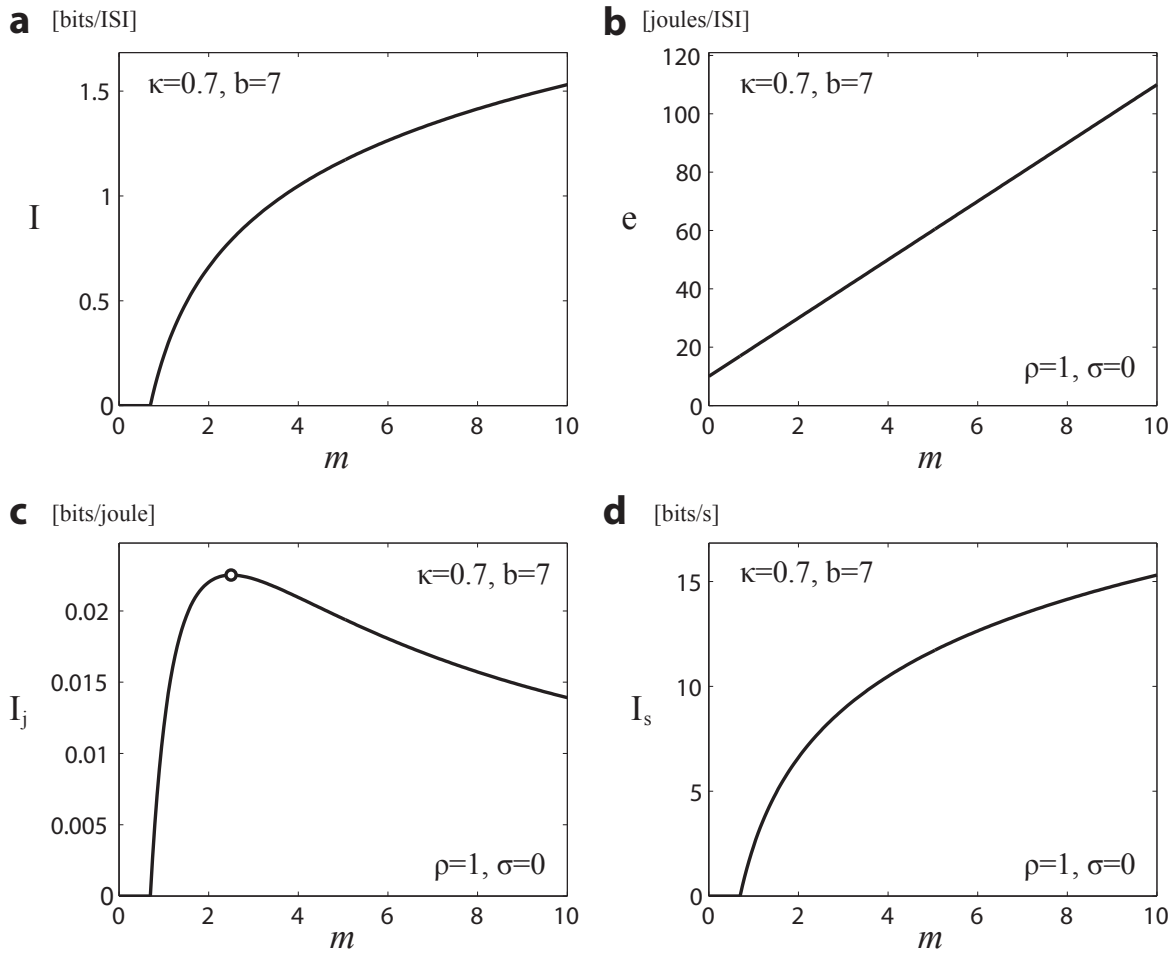

Figure 5.4: Best fit parameters of model to data for optimizing information/energy.

5.4 Methods Summary

- Electrophysiology. In vivo patch recordings, in cell-attached or whole-cell mode, were made from adult cats anesthetized with propofol and sufenta [55] [56] [57] [58] and were digitized at 10 or 25 kHz.

- Theoretical analysis. Based on the same assumptions used in the Berger-Levy energy efficient neuron model [5] [6], the optimal input and output probability density functions have been redefined for synaptic transmission. The closed-form solutions for mutual information and energy expenditure have been derived analytically.

- Fitting. Parameters of the bits-per-joule optimal input and output distributions of the Berger-Levy model were fitted to the statistics of RGC and LGN spike trains obtained from recordings [55]. The postsynaptic thalamic ISI distribution contains a scale parameter b and a shape parameter κ. The presynaptic retinal EPSP averaged excitation intensity distribution features three parameters, b, κ and m. Parameters for the Berger Levy model were fitted by a maximum likelihood (ML) estimation and average spiking ratio between RGC and LGN. The confidence interval was set at 95%.

5.4.1 Theoretical Analysis

To quantify the information transferred and energy expended during retinothalamic transmission, we applied the Berger-Levy energy efficient neuron model to a single input case. Four assumptions are adopted from the model: (1) The number of spike arrivals is governed by an inhomogeneous Poisson process; (2) Integrate-and-fire neuron with a threshold m is used as the basic computational model; (3) Mean value assumption; (4) Afferent inputs are excitatory and have equal synaptic weights.

For synaptic transmission, in order to achieve maximized information transferred per unit energy expended, (1) The postsynaptic ISI has a gamma distribution with parameters κ and b,

$$f_T(t) = \frac{b^\kappa t^{\kappa-1} e^{-bt}}{\Gamma(\kappa)} \tag{5.3}$$

(2) The scaled reciprocal of presynaptic averaged firing rate is beta distributed with parameters κ, b, and m,

$$f_\Lambda(\lambda) = \begin{cases} \frac{\Gamma(m)}{\Gamma(\kappa)\Gamma(m-\kappa)} \cdot \frac{b^\kappa (\lambda-b)^{m-1-\kappa}}{\lambda^m}, & \lambda \geq b \\ 0, & \lambda < b \end{cases} \tag{5.4}$$

Then the information and energy terms are defined as:

(3) The Shannon mutual information between averaged input firing rate Λ and output ISI duration T,

$$I(\Lambda;T) = \int_0^\infty \mathrm{d}\lambda f_\Lambda(\lambda) \int_0^\infty \mathrm{d}t f_{T|\Lambda}(t|\lambda) \log \frac{f_{T|\Lambda}(t|\lambda)}{f_T(t)} = O(\log m). \tag{5.5}$$

Omitting unwieldy mathematical derivations, the mutual information finally follows

$$I(\Lambda;T) = (\log\Gamma(\kappa) - \kappa\psi(\kappa) + \kappa) - (\log\Gamma(m) - m\psi(m) + m), \tag{5.6}$$

where $\Gamma(\kappa)$ denotes the gamma function and $\Psi(\kappa)$ represents the digamma function.

(4) The energy expenditure Z is defined as a sum of AP generation cost C_0, postsynaptic accumulation cost C_1M and basal metabolic cost C_2T,

$$Z = C_0 + C_1M + C_2T$$

.

Henceforth, after taking expectation of Z over both M and T it follows that

$$E[Z] = C_0 + C_1\mathbb{E}[M] + C_2\mathbb{E}[T] \tag{5.7}$$

$$= C_0 + C_1m + C_2\kappa/b, \tag{5.8}$$

where the expected value of random EPSP arrivals M is defined as m and the expected value of random ISI length T equals κ/b due to the fact that even when the synaptic weights are not required to be equal to one another, if the performance criterion remains maximization of bits per joule, then the output ISI durations remain gamma distributed with the same parameters κ and b as in the special case in which all synapses are assumed to have the same

weight.

Therefore after defining ρ as C_1/C_0, σ as $\frac{\kappa C_2}{bC_0}$, C_0 as a normalizing factor C and dividing C on both sides, the expected unit energy cost e can be written as

$$e = \mathbb{E}[Z]/C = 1 + \rho m + \sigma = O(m). \tag{5.9}$$

A more refined version of B-L theory has been developed in which membrane potential within each ISI accumulates like a Wiener process with drift. In that version the gamma distribution is replaced by a generalized inverse Gaussian probability density function.

5.4.2 LGN Neuron Model

The Berger-Levy theory of energy efficiency was based on an integrate-and-fire model neuron, which did not include variable amplitudes or decay of the excitatory postsynaptic potential (EPSP). Although this model neuron fits experimental data in the medium and large firing rate domain, it does not perform well for low average firing rates (Fig. 5.5). In order to address this discrepancy, we modified the neuron model to account for the phenomenon of membrane potential decay in LGN cells during EPSP accumulation. The goal of this model is to measure and estimate the EPSP accumulation process with higher precision, enabling comparison between estimated spike timing statistics and experimental data.

The EPSP accumulating process for the leaky integrate-and-fire model neuron is illustrated in Figure 5.4b. The model neuron had four parameters:

- β: the exponential decay rate
- θ: the action potential (AP) firing threshold
- h: the resting membrane potential reset
- μ: the amplitude of the incoming EPSP.

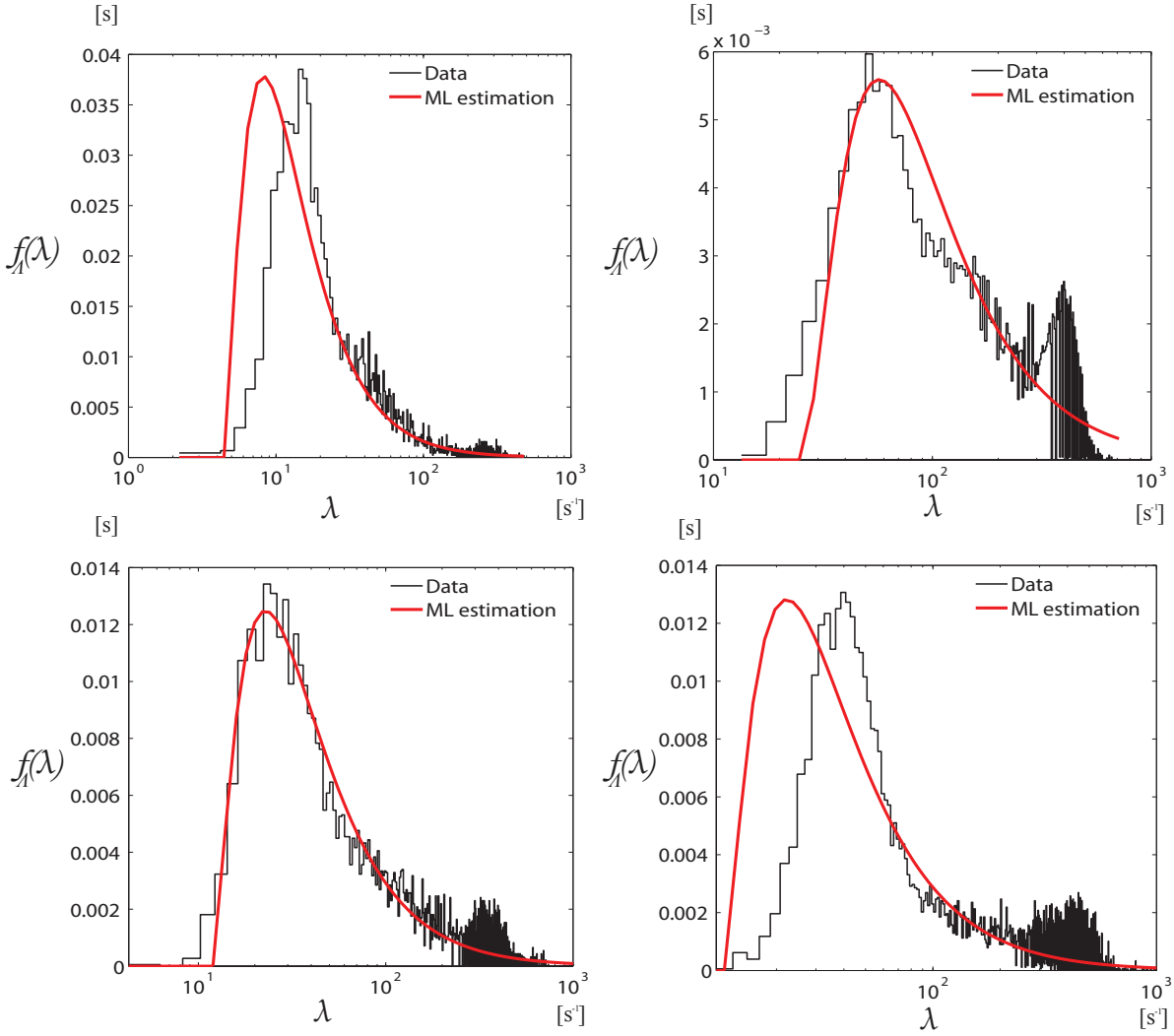

Figure 5.5: Retinal averaged firing rate parameter estimation from four typical data sets.

First, after each interspike interval (ISI) the membrane potential is reset to h after an absolute refractory period Δ. When each EPSP arrives at the spike initiating zone, it will immediately make a contribution by a unit step function with amplitude μ and then decay exponentially at rate β. Hence, after receiving the first EPSP arrival during the kth LGN ISI, the membrane potential becomes

$$V_{EPSP_1} = he^{-\beta\Delta_{t_{EPSP_1}}} + \mu, \tag{5.10}$$

where $\Delta_{t_{EPSP_1}} := t_{EPSP_1} - (t_{Spike} + \Delta)$.

Likewise, V_{EPSP_2} turns out to be

$$V_{EPSP_2} = V_{EPSP_1} e^{-\beta \Delta_{t_{EPSP_2}}} + \mu, \tag{5.11}$$

where $\Delta_{t_{EPSP_2}} := t_{EPSP_2} - t_{EPSP_1}$. Therefore, when m EPSPs arrive during the kth ISI, the accumulated membrane potential V_{EPSP_m} is

$$V_{EPSP_m} = V_{EPSP_{m-1}} e^{-\beta \Delta_{t_{EPSP_m}}} + \mu, \ \forall m \geq 2, \tag{5.12}$$

where $\Delta_{t_{EPSP_m}} := t_{EPSP_m} - t_{EPSP_{m-1}}$. Hence, we must have both $V_{EPSP_k} < \theta$ for $k = 1, \cdots, m-1$ and $V_{EPSP_m} \geq \theta$ in order to trigger an AP.

A flow chart of the computational procedure used to fit the spike timing data is shown in Figure 5.6. The histogram of the accumulated membrane potential after selecting certain typical values for the parameters in Figure 5.6b is asymmetric when the decay rate β is small. We set the threshold θ to best capture the experimental data so that the total of the magnitudes of the differences between the true LGN neuron's actual spiking time and the model's estimate thereof is minimized. The magnitude difference is easy to calculate in those ISI's in which the model spikes at an earlier time than when the actual LGN neuron does. However, when the model fails to reach the threshold before the data does, there is no more EPSP data for the terminated ISI. Hence it is not directly possible to determine how late the model would have fired a spike. We could, of course, take the EPSP arrivals in the subsequent ISI(s) until the model spikes, but then we would be double-using some EPSP's arrival times. Instead of doing that, whenever the model was late in spiking relative to the data, we generated exponential random variables using the long-term statistics of the input EPSP train until the model crossed the threshold (Fig. 5.6a).

This procedure optimizes the timing error instead of the voltage error. The algorithm for computational thresholding (Fig. 5.6a) converts the voltage error into an approximate

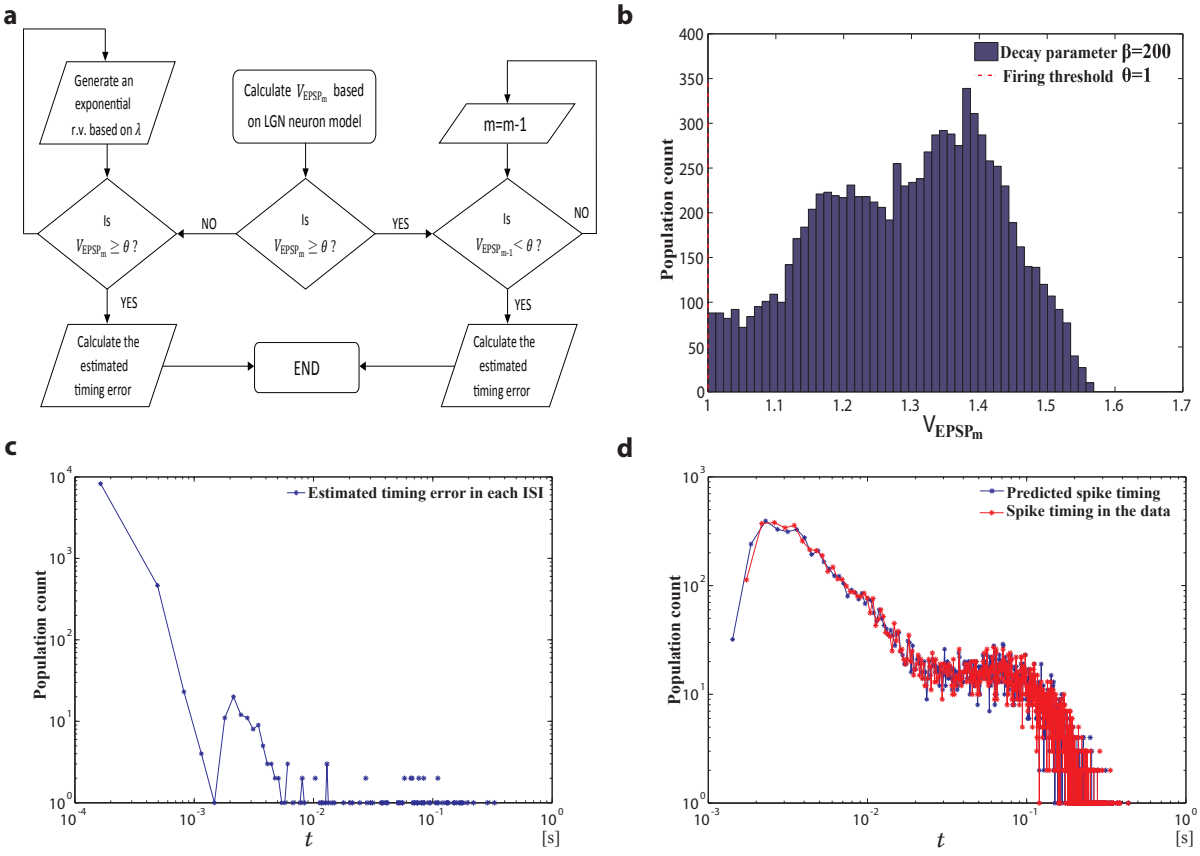

Figure 5.6: Detailed thresholding algorithm and computational simulation results.

timing error. After considering all feasible combinations of the parameters, we found for each data set a value (or values) of the (β, θ, h, μ) quadruple that minimized the absolute value of the estimated timing error. The values of the parameters for one typical data set are: $(\beta, \theta, h, \mu) = (220, 1, 0.43, 0.57)$. The mean of the minimized timing error is 0.928 ms and the standard deviation (SD) of the minimized timing error is 9.1 ms. Not surprisingly, a few outliers are responsible for much of this big absolute error SD. If one is willing to disregard 0.16% of the 8916 ISIs in this data set, then the SD is reduced by more than 40%. Figure 5.6c shows the histogram of the estimated timing error. Figure 5.6d compares the estimated timing error with the AP timing statistics in the experimental data. Moreover, gamma

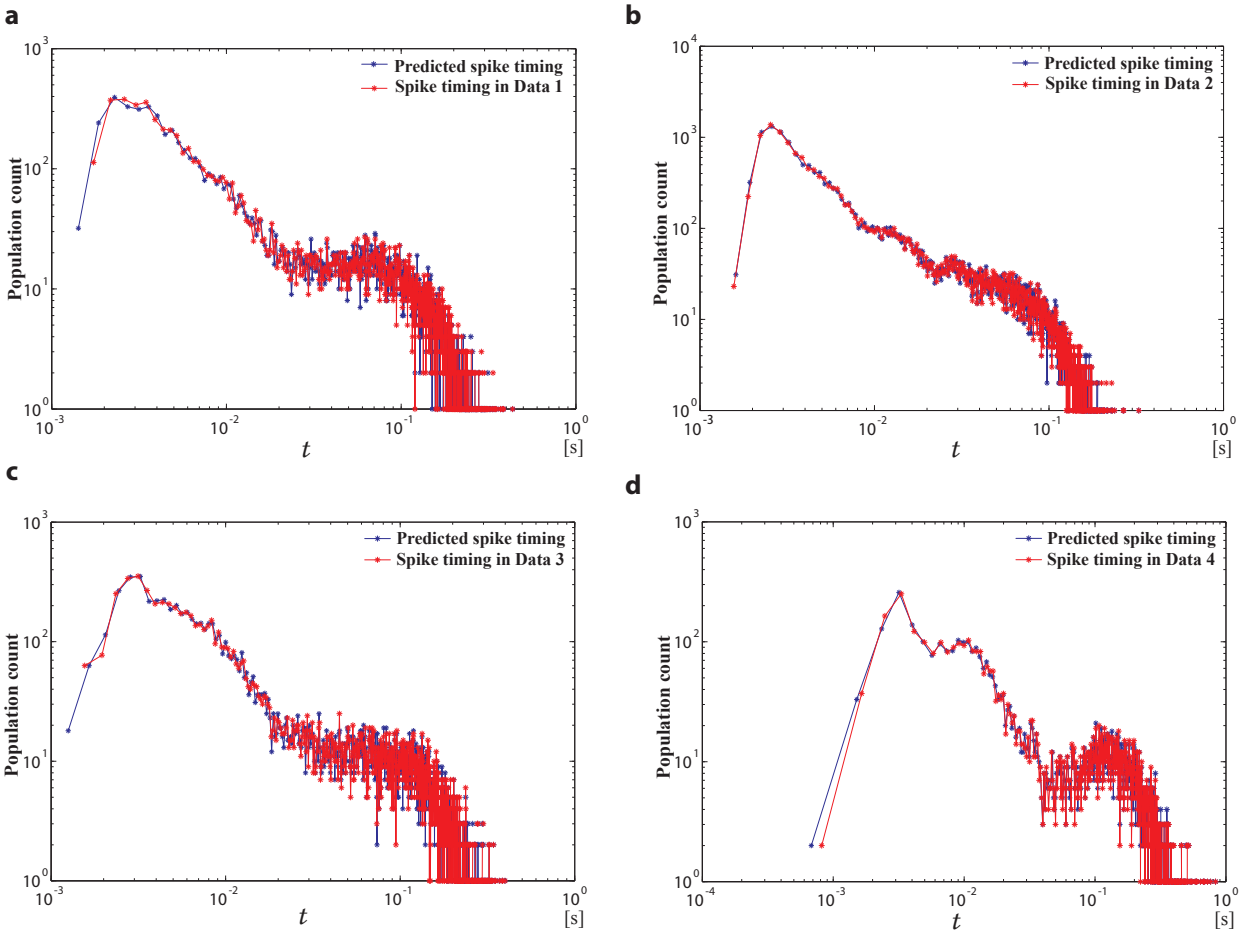

Figure 5.7: Thalamic spike timing estimated from four typical data sets. The histograms of spike timing (blue) are compared with the predicted ones from the LGN neuron model (red).

distributed output LGN ISI statistics are obtained when the input EPSP timings are fed into the computational LGN neuron model as is predicted by the information-energy optimization inherent in Berger-Levy theory. The result of applying this procedure to recordings from 4 typical retinal neurons is shown in Figure 5.7.

5.4.3 Two-to-four Fold Reasoning

The B-L theory predicts input and output probability distributions that maximize the bits of information a neuron conveys to its efferent targets per joule of energy it expends for postsynaptic accumulation, action potential (AP) propagation and vesicle restoration,

clocking, and basal metabolism. The theory implies that (1) the intervals between consecutive output spikes must be distributed according to a gamma distribution with shape and scale parameters denoted by κ and b, respectively, and (2) the normalized probability density of the scaled reciprocal of the neuron's net afferent excitatory and inhibitory intensity must be distributed according to a beta distribution with three free parameters, two of them being the aforementioned κ and b and the third being a threshold-related parameter, m. In this Chapter, m has been extended to be the ML estimate of a parameter thereby resulting in m usually being a non-integer.

In the data analysis, specific values for κ and b have been obtained by matching the empirically obtained output interspike interval (ISI) distribution with the theoretically derived gamma distribution according to the maximum likelihood (ML) method. Then, given fixed κ and b, m is chosen to make the empirical distribution of the scaled reciprocal input intensity best match a (κ, b, m) beta distribution.

In order to obtain the resulting bits per joule performance predicted by the B-L theory, one not only needs marginal distributions of input and output that conform well to the B-L theory, but also needs the channel kernel (that is, the conditional distribution $f_(T|\lambda)(t|\lambda)$) to conform to the mathematical prescription specified in B-L theory, extended by letting m be any positive real number no longer required to be an integer. In theory, the information rate is a functional of the joint distribution of T and Λ, which turns out to be a function of only κ and m. The energy expenditure, however, is defined as a function of m, ρ and σ. Here, ρ represents the normalized energy cost ratio between processing a retinal spike and producing a LGN spike, and σ represents the normalized metabolic energy cost during one ISI.

The optimization of information rate subject to energy constraints can be reduced to a concave maximization problem with respect to the single parameter m provided the channel kernel matches well with the neural data and that the parameters κ is chosen appropriately. That is, there is always an optimal value for m that maximizes information transmitted per joule expended. Moreover, with specific values of ρ and σ chosen, the information-energy

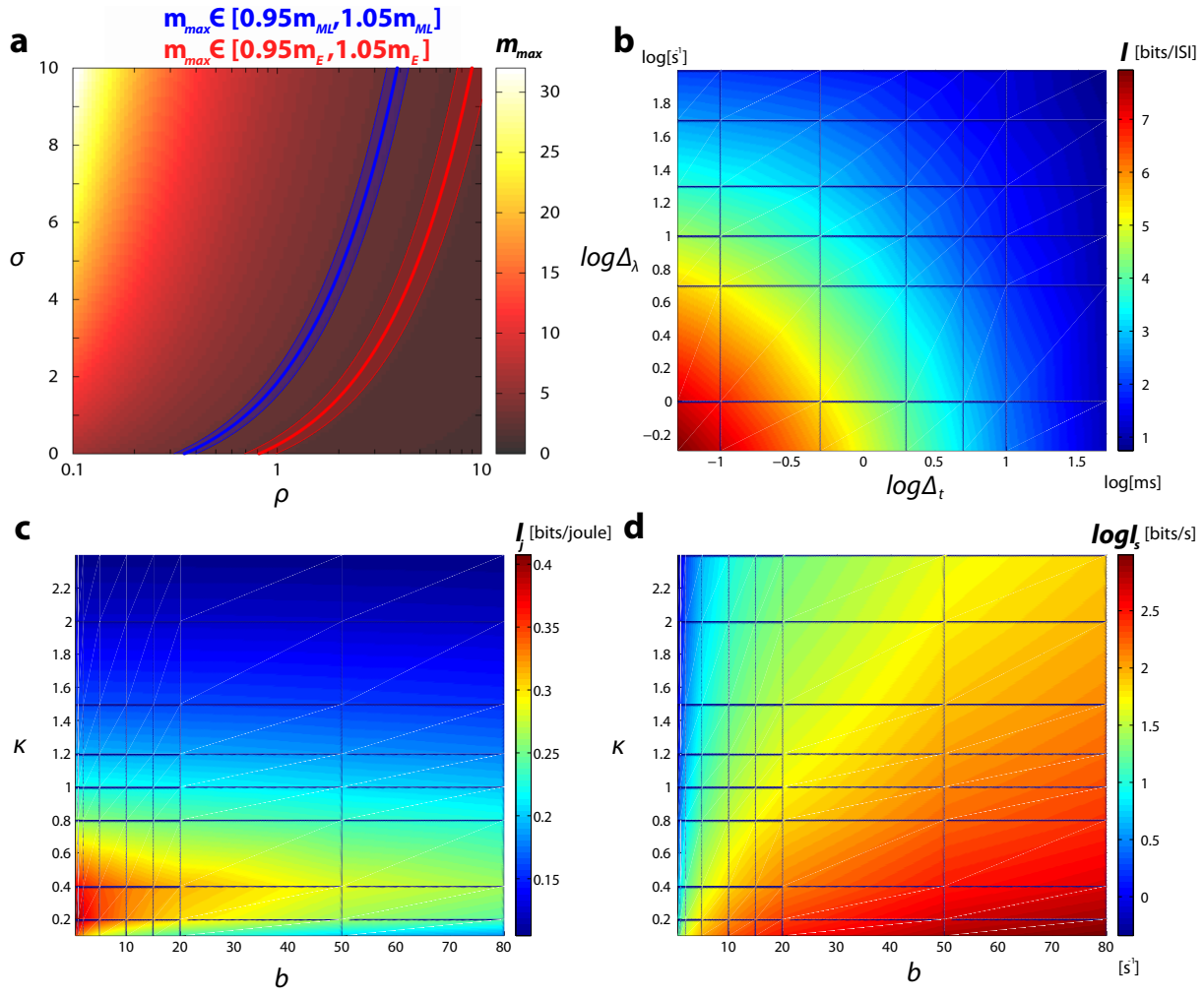

Figure 5.8: Parametric spaces of bits per joule optimized average number of arrivals per ISI (m_{max}), mutual information estimation (I), information per energy (I_j), and information per second (I_s).

optimized value of m can be tuned into a 2- to 4- regime.

Unfortunately, the recorded data are too sparse to estimate the conditional distribution accurately. Hence, we have to assume that the analytical form of the channel kernel specified by extended BL theory would be well-matched by more extensive recordings from the LGN neuron.

Figure 5.8 indicates parametric spaces of bits per joule optimized average number of arrivals per ISI (m_{max}), mutual information estimation (I), information per energy (I_j), and

information per second (I_s). A particular case for specific ρ and σ values is illustrated in Figure 5.4, where m is between 2 to 4. In order to conclude that the BL theory is a good explanation for the RGC-LGN spike ratio, one needs to have additional information about the values of the energy ratio ρ and the metabolic cost σ. Exploring the data sets, we have observed that the average number of RGC spikes per ISI, m, ranges from 2 to 4. Before one can assess the goodness of the fit and hence the degree of validity of the BL theory prediction, one has to obtain measurements of ρ and σ. In order to determine the energy expended by the LGN neuron, one needs to count both (1) the energy expended to process not only RGC spikes but also spikes afferent to other synapses of the LGN neuron with neurons in the visual hierarchy, and (2) the energy expended to propagate the LGN neuron's spikes to all its targets and to restore the neurotransmitter vesicles inside the presynaptic terminals at said targets. Unfortunately, current neuroscience technology does not permit the acquisition of such measurements.

5.4.4 Mutual Information Estimation of the Spike Trains

The direct binning method is employed in order to estimate the mutual information from the experimental spike train. We first derive the formula for mutual information using finite linear bins. According to the definition of mutual information,

$$I(\Lambda;T) = \sum_{\lambda}\sum_{t} \Pr(\lambda,t)\log_2 \frac{\Pr(\lambda,t)}{\Pr(\lambda)\Pr(t)} \tag{5.13}$$

$$= \sum_{\lambda}\sum_{t} f_{\Lambda,T}(\lambda,t)\Delta_\lambda\Delta_t \log_2 \frac{f_{\Lambda,T}(\lambda,t)}{f_\Lambda(\lambda)f_T(t)} \tag{5.14}$$

Based on the sampling frequency in the experimental recordings, the mutual information can be calculated as 2.72 bits when picking the resolutions $\Delta_\lambda = 10$ spikes/s and $\Delta_t = 50$ ms.

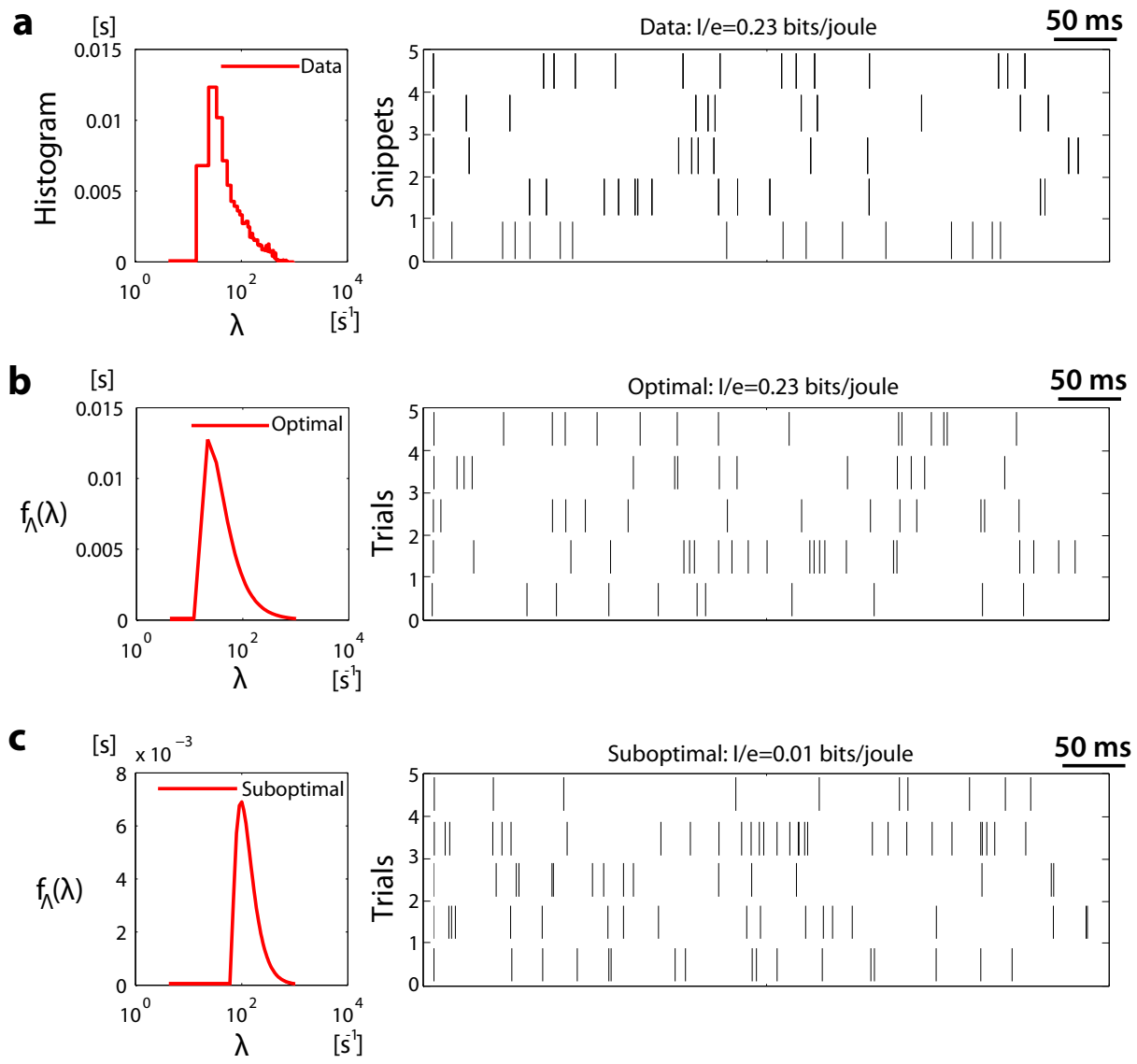

Figure 5.9: *In vivo* and *in silico* optimal and suboptimal spike trains demonstration.

5.4.5 Optimal and Suboptimal Spike Trains Demonstration

By maximizing the average mutual information rate over a constraint on the total energy cost that a neuron expends for metabolism, postsynaptic potential generation, and action potential propagation during one ISI, Berger-Levy theory obtain the rescaled reciprocal of the average firing rate as a beta distribution functional with parameters, m, b and κ. Sub-optimality can be achieved by changing the spike train pattern such that the average

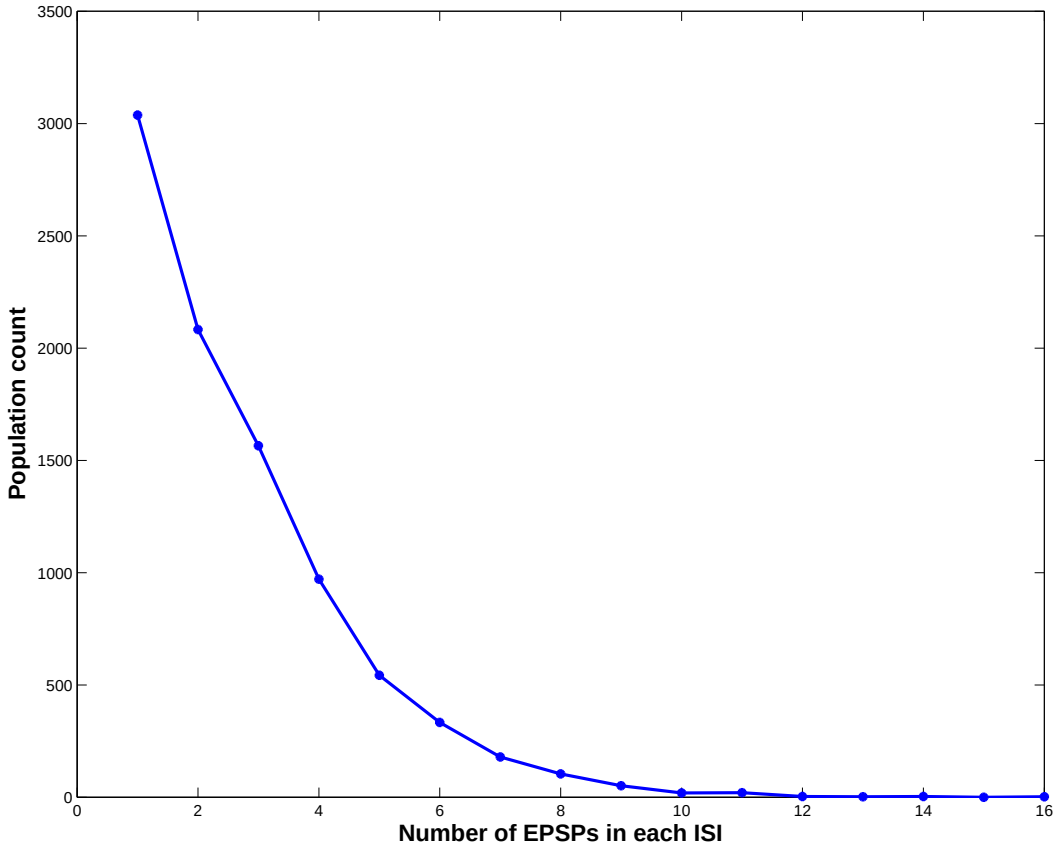

Figure 5.10: Histogram of the number of RGC arrivals in each LGN ISI from one typical data set.

mutual information rate is held fixed but average energy expenditure is increased, or vice versa. In other words, in order to be suboptimal one can either change the distribution form of the optimized functional or adjusting the corresponding parameters, m, b and κ. In vivo and in silico optimal and suboptimal spike trains have been demonstrated in Figure 5.9 with information/energy values calculated (red).

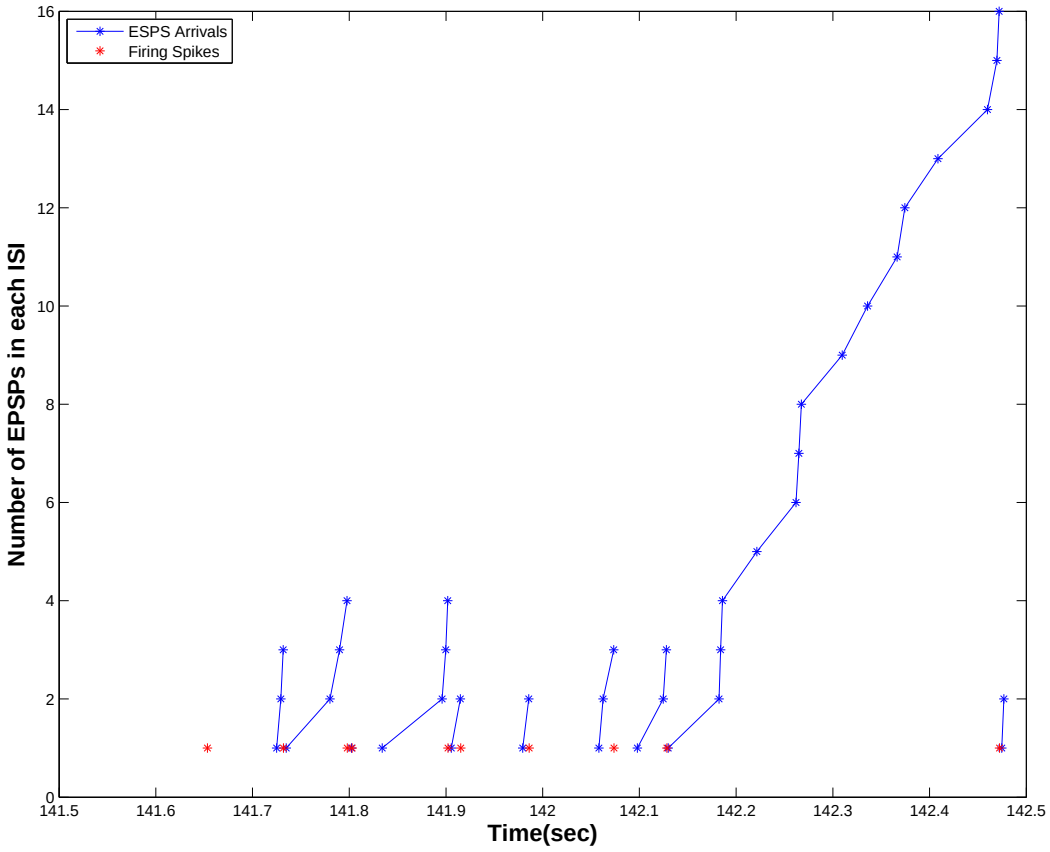

Figure 5.11: Demonstration of arriving EPSPs in consecutive ISIs.

5.4.6 Power Law Distributed RGC Arrival Numbers in Each LGN ISI

We obtain the histogram of the EPSP arriving numbers for each LGN ISI from a particular RGC/LGN data set (Fig. 5.10). As shown in Figure 5.11, there can be up to 16 arrivals for this typical data set even though the average value is around 2.65, which agrees with our expectation of 2 to 4. Moreover, the histogram exhibits a power law distribution, which has been verified from other data sets as well. According to Figure 5.10, in more than half of the LGN ISIs a spike is fired before a third input can arrive. However, if it doesn't fire from the first or the first/second or the first/second/third combined inputs, then occasionally two but

usually three or more arrivals close together are required in order to make the accumulated postsynaptic potential cross the threshold. Our leaky LGN neuron model covers all these modes of firing. These firings on the first/second incoming spikes are the mechanism that LGN uses to convey to its targets that its inputs are arriving at a rate that is considerably above average.

5.4.7 Comparison between Two Point Estimators of Parameter m

For a parameter of interest in a model, such as m in the B-L energy efficient theory, the objective of point estimation is to use a sample to compute a number that represents a good guess for the true value of the parameter. Generally speaking, there are two methods of obtaining point estimates for m: the method of moments, m_E, and the method of maximum likelihood, m_{ML}. For the method of moments, the basic idea is to equate certain sample characteristics, such as the mean, to the corresponding physical parameter in the model in order to yield the estimator. For the maximum likelihood estimator, let the random variables $X_1, X_2, \cdots, X_n$ have joint pmf or pdf $f(X_1, X_2, , X_n; \theta_1, \theta_2, \cdots, \theta_l)$, where the parameters $\theta_1, \theta_2, \cdots, \theta_l$ have unknown values. When $x_1, x_2, \cdots, x_n$ are observed data sample values, $f(x_1, x_2, \cdots, x_n; \theta_1, \theta_2, \cdots, \theta_l)$ is regarded as a function of $\theta_1, \theta_2, \cdots, \theta_l$ and becomes the likelihood function. The maximum likelihood estimates (mle's) $\theta_1, \theta_2, \cdots, \theta_l$ are those values of the θ_i's that maximize the likelihood function so that $f(x_1, x_2, \cdots, x_n; \theta_1, \theta_2, \cdots, \theta_l) \geq f(x_1, x_2, \cdots, x_n; \theta_1, \theta_2, \cdots, \theta_l)$ for all $\theta_1, \theta_2, \cdots, \theta_l$. When the sample size n is large, the maximum likelihood estimator of any parameter θ_i is approximately unbiased and has variance that is nearly as small as can be achieved by any estimator, i.e., the mle (θ_i) is approximately the minimum variance unbiased estimator (MVUE) of θ_i. In this specific analysis, $l = 3$ with $\theta_1 = b$, $\theta_2 = \kappa$, $\theta_3 = m$. b and κ have already been set at ML estimate of the parameters of the gamma distribution for the pdf of T. So m is the only remaining parameter that remains for which the ML estimate has not yet been determined for the pdf of Λ.

For a typical data set, we are able to calculate three different values for the average number of RGC arrivals within each LGN ISI duration.

- Experimental data: the average number of arrivals in each ISI is $m_E = 2.65$.
- B-L energy efficient integrate-and-fire (IF) neuron model: $m_{ML} = 3.01$ for $b = 13.46$ and $\kappa = 0.67$.
- Computational leaky integrate-and-fire (LIF) LGN neuron model: $m_{Leaky} = 3.43$.

The main reason why m_{Leaky} is larger than m_{ML} and m_E is that on those occasions where the data fired before the leaky model did, the subsequent RGC arrival times were independently and identically selected as opposed to the fact that the data shows that an RGC ISI that is considerably shorter than average tends to be followed by another such ISI, thereby allowing less leakage and hence usually reaching the threshold with fewer RGC arrivals.

Since a point estimate by itself provides no information about the precision and reliability of the parameter estimation, a 95% large-sample confidence interval (CI) has been calculated for m as:

$$(2.65 - 1.96 \times 1.83/\sqrt{8916}, 2.65 + 1.96 \times 1.83/\sqrt{8916}) \approx (2.61, 2.68).$$

5.4.8 Detailed Figure Captions

- Figure 5.1: Retinothalamic transmission. **a**, Schematic diagram of retinal ganglion cell (blue) synapsing on a thalamic relay cell in the LGN (red), which in turn projects to the cat visual cortex. **b**, In vivo "cell-attached" recording from LGN relay cells. **c**, Recordings of retinal (blue) and thalamic (red) spike trains.
- Figure 5.2: LGN neuron model. **a**, The afferent inputs (blue) generate EPSPs in the LGN neuron while the efferent outputs (red) correspond to the generated APs. The LGN neuron model used to fit the ISI distribution is a leaky integrate-and-fire

neuron with four parameters (β, θ, h, μ), where θ is the firing threshold, h is the level to which the membrane potential is reset after an absolute refractory period Δ after each spike, and μ is the amplitude of an EPSP which decays exponentially at rate β (see supplementary information). **b**, the sequence $\{T_i\}$ is a random process representing the output ISI durations while the sequence $\{\Lambda_i\}$ is a random process featuring the averaged firing rate of the input EPSPs and i represents the ISI index. V_m stands for the postsynaptically accumulated potential. M is the number of arrivals needed within one ISI to cross the threshold. **c**, Illustration of an instance in which five EPSPs accumulate within one ISI. The membrane potential V_m integrates the postsynaptically integrated EPSPs.

- Figure 5.3: Thalamic interspike interval (ISI) parameter estimation and retinal averaged firing rate parameter estimation. **a**, Histogram of a typical postsynaptic thalamic cell's ISI (black), and probability density function of a gamma distribution with a typical parameter set, κ and b, (red). **b**, Parametric space, κ and b, for the population results of X-cells (purple), Y-cells (green), and unidentified cells (black) with median value (red). **c**, Histogram of a typical presynaptic retinal ganglion cell's averaged firing rate (black), and probability density function of a scaled reciprocal of beta distribution with the same parameters, κ, b, and ML estimated m (red). **d**, Parametric space, m_{ML} and m_E, for the population results of presynaptic X-cells (purple), Y-cells (green), and unidentified cells (black) with median value (red).

- Figure 5.4: Best fit parameters of model to data for optimizing information/energy. **a**, Information (I) plotted as a function of m when $\kappa = 0.7$. **b**, Energy (e) plot as a function of m when $\kappa = 0.7$, $b = 7$, $\rho = 1$ and $\sigma = 0$. **c**, Bits per joule (I_j) plot as a function of m when $\kappa = 0.7$, $b = 7$, $\rho = 1$ and $\sigma = 0$, with the maximal value indicated by an open circle. **d**, Bits per second (I_s) plot as a function of m when $\kappa = 0.7$, $b = 7$, $\rho = 1$ and $\sigma = 0$.

- Figure 5.6: Detailed thresholding algorithm and computational simulation results. **a**, Computational thresholding algorithm. **b**, Accumulated EPSP value, V_{EPSP_m}, in each ISI with a fixed threshold. **c**, Estimated timing error histogram. **d**, Output ISI duration comparison between the model (Blue) and the data (Red).

- Figure 5.8: Parametric spaces of bits per joule optimized average number of arrivals per ISI (m_{max}), mutual information estimation (I), information per energy (I_j), and information per second (I_s). **a**, Parametric space, σ and ρ, for the energy expenditure per ISI showing the 95% confidence intervals for maximum likelihood value of m_{ML} (blue) and expected value m_E (red) that optimizes information/energy gives $m_{max} = 2 - 4$ (color bar). **b**, Mutual information, I, estimation with respect to varying resolution in terms of ISI duration, Δ_t, and averaged firing rate, Δ_λ. **c**, Parametric space of bits per joule, I_j, with respect to κ and b. **d**, Parametric space of bits per second, I_j, with respect to κ and b.

- Figure 5.9: *In vivo* and *in silico* optimal and suboptimal spike trains demonstration. Experimental data is shown in (**a**). Theoretical predictions under the optimal condition are illustrated in (**b**). The suboptimal scenarios are demonstrated in (**c**). Panels on the left represent the averaged firing rate statistics per ISI; Panels on the right represent the recorded (**a**) with five snippets and simulated (**b**, **c**) spike trains with five trials.

Chapter 6

Conclusions and Future Work

Everything but our understanding is flawless.

-Archie Randolph Ammons

6.1 Conclusions

In this thesis, we have shown that, when neuron j is designed to maximize bits conveyed per joule expended and employs a mixture of Gamma distribution as the channel density function, even though j's synapses are no longer being required to all have the same weight or to be excitatory only, the pdf of the ISI durations continues to be a delayed gamma distribution as it was in [5] wherein all the weights were assumed to be equal. This happens despite the fact that the conditional distribution for T given Λ is now a mixture of gamma distributions instead of the pure gamma distribution that characterizes the special case of equal weights.

Additionally, we have implicitly determined the optimal distribution $f_\Lambda(\lambda)$ that characterizes the afferent excitation/inhibition intensity by (1) maximizing the Shannon mutual information rate given a constraint on the total energy cost (a cortical neuron's energy expenditure for metabolism, postsynaptic potential accumulation, and action potential generation and propagation during one ISI); (2) converting the integral equation to a differential equation with a closed-form solution.

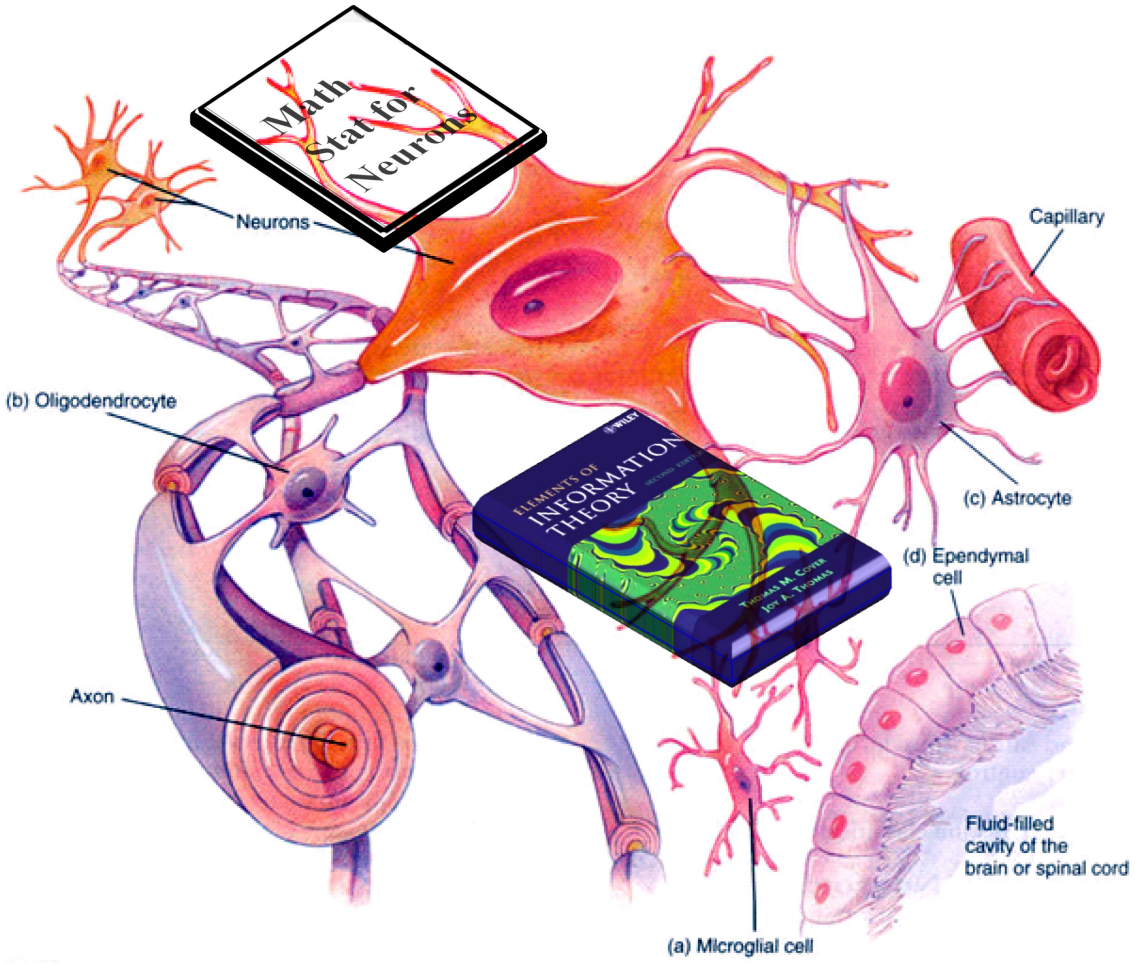

Figure 6.1: Illustration of neurons learning mathematical statistics and information theory.

Moreover, we have also shown that, when neuron j employs a IG/GIG conditional distribution as the channel density function and when the bpJ-maximization is achieved, the output ISI distribution is a related GIG marginal distribution. This has allowed us to compute the tradeoff between the information rate and the average power in the IG/GIG class in a markedly simpler way requiring only a one dimensional integral instead of several multidimensional integrals. Furthermore, we have obtained the associated input distribution $f_\Lambda(\lambda)$ in the IG case and shown how to numerically obtain the associated input distribution $f_\Lambda(\lambda)$ in the GIG case. By generalizing from the Gamma and IG families to the GIG family, the derived results contain [5] [6] [7] [8] as special cases in which the three parameters are set

to specific values.

Last but not least, we have employed our energy efficient theory to resolve the paradoxical energy efficiency of retinothalamic neural transmission. By imposing an energy function that combines metabolism, PSP accumulation and LGN spiking costs during each ISI, the theory predicts that, in order to achieve the highest amount of transmitted information per unit energy expenditure, the number of presynaptic spikes should be approximately 2 to 4 times greater than the number of thalamic spikes, consistent with our experimental observations from anesthetized cats. This is the first time that a quantitative experimental and theoretical analysis has been performed on information transmission subject to an energy constraint in an isolated nucleus. Our finding suggests that the brain does not optimize information flow only, but rather energy-efficient information flow.

The energy efficiency of the human brain in terms of information processing in problems of pattern recognition and optimized adaptive decision making is astonishingly superior to that of man-made machines. By extending the information-energy efficient neuron model to a more general framework including both unequal synaptic weights and inhibitory synapses, our theory closely corresponds with the actual neurophysiology of cortical networks, potentially leading to wider applications in neuroscience and engineering.

6.2 Future Work

6.2.1 KKT instead of Lagrange Multipliers

Equation (4.56) specifies the exact expression for the bpJ-maximizing probability density of the random excitation intensity, Λ. There are two significant aspects of this equation. One is that it extends the solution for the bpJ-maximizing afferent intensity's pdf from the idealized IIF neuron model treated by Berger and Levy in 2010 to the far more neuroscientifically relevant family of GIG neuron channel models. The other is that it shows that such a probablilty density exists only when the parameter vectors (α, β, γ) and (a, b, c) are properly

related. Specifically, if the integral in this equation is finite for all values of $\lambda > 0$, then λ is an absolutely continuous random variable the pdf of which is given by said equation. However, it is easy to see that if the arithmetic values of the two vectors are exchanged (i.e., if $\alpha \leftrightarrow a$, $\beta \leftrightarrow b$ and $\gamma \leftrightarrow c$), then if the integral converges before these exchanges are made, then it will not converge after these exchanges. This means that sometimes Λ has a pdf and sometimes it doesn't, depending on the values of the two parameter vectors. When the integral does not converge, our Lagrangian solution is not correct and the optimizing distribution for Λ does not possess a pdf. This is why it is necessary to study how KKT theory could be used to treat these situations in which Λ is not absolutely continuous. We believe that these situations correspond to cases in which the energy constraints are too stringent for the neuron's current design and conjecture that this situation occurs only rarely in practice. However, the most significant instances of natural selection usually take place when such imbalances occur. Therefore, it is important to try to apply KKT theory to get a more robust solution.

6.2.2 What Do Neurons Do and Not Do Highly Energy Efficiently?

Our current belief is that neurons pinpoint highly energy efficiently the exact times at which an AP occurs and the exact times at which an AP arrives at a synapse. This is done in the first case by means of fast sodium channels that reside at the initial segment of the axon and in the second case by fast calcium channels located at the presynaptic terminals. Our present conjecture is that these two operations occur with close to $kT \log 2$ Joules being expended per bit of information encoded in the first case and decoded in the second case. We also believe at present that the dendro-somatic processing of EPSP's and IPSP's does not occur nearly that close to the thermodynamic limit but more like a factor of 10 or possibly even 20 times said limit. What the brain gains for this thermodynamic non-optimality is the ability to have a sophisticated network of billions of neurons that need to expend energy communicating with one another, something that is hard to do energy efficiently because of its intrinsically

large spatial extent. Apparently, the capabilities that advanced organisms gain from this network must justify the reduction in energy efficiency. For example, humans have found out many things by means of the enormous in-degrees and out-degrees of connectivity among the neurons in their brains, including the fact that a vast amount of the Earth's energy is stored way underground! No other animal knows that. This has proved to be "energy efficient" in a certain sense, although it is becoming increasingly clear that we are abusing this knowledge from the viewpoint of efficient use of that energy.

6.2.3 Network Coding

Network coding research has convincingly shown that store-and-forward message passing can be highly inefficient. Instead of just storing and forwarding packets as is presently done in the Internet, it sometimes can be much more efficient to implement clever coding at each node. The brain appears to do this in spades, doubled and redoubled, when one considers that each cortical neuron simultaneously receives thousands of spike train inputs from other cortical neurons but puts out only a single spike train of its own; however, it delivers that spike train to thousands of other neurons. Thus, it appears that brains do network coding in the extreme. We hope to find a way to quantify why this is a good way to behave even though it may not be as energy efficient as what small animals with few neurons are capable of achieving by virtue of not having to spend lots of energy on intra-organism communication.

Bibliography

[1] J. M. Kinney, H. N. Tucker, Clintec International Inc, *Energy Metabolism: Tissue Determinants and Cellular Corollaries (Raven, New York)*, p xvi, 1992.

[2] L. C. Aiello, P. Wheeler, *The expensive-tissue hypothesis–the brain and the digestive-system in human and primate evolution*, Curr Anthropol 36:199-221,1995.

[3] D. Modha, *Private communication*, 2009.

[4] A. Hasenstaub, S. Otte, E. Callaway, and T. J. Sejnowski, *Metabolic cost as a unifying principle governing neuronal biophysics*, Proc Natl Acad Sci USA 107: 12329-12334, 2010.

[5] T. Berger and W. B. Levy, *A mathematical theory of energy efficient neural computation and communication*, IEEE Trans. IT, vol. 56, No. 2, pp. 852-874, February 2010.

[6] J. Xing, T. Berger, T. J. Sejnowski, *A Berger-Levy Energy Efficient Neuron Model with Unequal Synaptic Weights*, 2012 IEEE International Symposium on Information Theory, pp. 2964-2968, Cambridge, MA, 2012.

[7] J. Xing, T. Berger, T. J. Sejnowski, *An energy efficient neuron model with unequal synaptic weights*, submitted to IEEE Trans. IT, December 16, 2013.

[8] T. Berger, W. B. Levy, J. Xing, J, *Energy efficient neurons with generalized inverse Gaussian interspike interval durations*, 2011 49th Annual Allerton Conference on Communication, Control, and Computing, 1737-1742, 2011.

[9] T. Berger, J. Xing, W. B. Levy, *Generalized inverse Gaussian (GIG) models for energy-efficient neurons*, Information Theory and Applications Workshop (ITA), pp. 1109, 2013, San Diego, CA.

[10] J. Xing, T. Berger, *Energy efficient neurons with generalized inverse Gaussian conditional and marginal hitting times*, 2013 IEEE International Symposium on Information Theory, pp. 1824-1828, 2013, Istanbul, Turkey.

[11] J. Xing, T. Berger, *Energy efficient neurons with generalized inverse Gaussian conditional and marginal hitting times*, submitted to IEEE Trans. IT, February 26, 2014.

[12] D. M. MacKay, W. S. McCulloch, *The limiting information capacity of a neuronal link*, Bulletin of Mathematical Biophysics, vol. 14, 127-135, 1952.

[13] W. S. McCulloch, *An upper bound on the informational capacity of a synapse*, In Proceedings of the 1952 ACM national meeting, Pittsburgh, Pennsylvania.

[14] R. R. de Ruyter van Steveninck, W. Bialek, *Real-time performance of a movement-sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences*, Proceedings of the Royal Society Series B, Biological Sciences, 234, 379-414, 1988.

[15] W. Bialek, F. R. Rieke, R. R. de Ruyter van Steveninck and D. Warland, *Reading a neural code*, Science, 252,1854-7, 1991.

[16] W. B. Levy, R. A. Baxter, *Energy efficient neural codes*, Neural Comput 8:531-543, 1996.

[17] W. B. Levy, R. A. Baxter, *Energy-efficient neuronal computation via quantal synaptic failures*, J Neurosci 22:4746-4755, 2001.

[18] V. Balasubramanian, D. Kimber, M. J. Berry, *Metabolically efficient information processing*, Neural Comput 13:799-815, 2001.

[19] B. Sengupta, M. Stemmler, S. B. Laughlin and J. E. Niven, *Action potential energy efficiency varies among neuron types in vertebrates and invertebrates*, PLoS Comput. Biol. 6, e1000840, 2010.

[20] J. J. Harris, R. Jolivet, D. Attwell, *Synaptic energy use and supply*, Neuron 75, 762-777, 2012.

[21] J. E. Niven, I. C. Anderson, S. B. Laughlin, *Fly photoreceptors demonstrate energyinformation trade-offs in neural coding*, PLoS Biol 5:e116, 2007.

[22] V. Balasubramanian, M. J. Berry, *A test of metabolically efficient coding in the retina*, Network 13:531-552, 2002.

[23] S. B. Laughlin, R. R. de Ruyter van Steveninck, J.C. Anderson, *The metabolic cost of neural information*, Nat Neurosci 1:36-41, 1998.

[24] D. Attwell, S. B. Laughlin, *An energy budget for signaling in the grey matter of the brain*, J Cereb Blood Flow Metab 21:1133-1145, 2001.

[25] J. J. Harris, D. Attwell, *The energetics of central nervous system white matter*, J. Neurosci. 2, 356-371, 2012.

[26] B. D. Willmore, J. A. Mazer, J. L. Gallant, *Sparse coding in striate and extrastriate visual cortex*, J Neurophysiol, 105(6):2907-19, 2011.

[27] S. B. Laughlin, T. J. Sejnowski, *Communication in neuronal networks*, Science 301:1870, 2003.

[28] US National Institutes of Health, National Institute on Aging created original - http://www.nia.nih.gov/Alzheimers/Publications/UnravelingtheMystery/.

[29] Analysis of neurotransmitter release mechanisms. Department of Genome Sciences, University of Washington.

http://www.gs.washington.edu/labs/pallanck/research/synaptic.htm Accessed on 15 January 2003.

[30] G. Mitchison, *Neuronal branching patterns and the economy of cortical wiring*, Proc Biol Sci 245:151-158, 1991.

[31] D. B. Chklovskii, A. A. Koulakov, *Maps in the brain: What can we learn from them*, Annu Rev Neurosci 27:369-392, 2004.

[32] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[33] R. W. Yeung, *Information Theory and Network Coding*, New York, Springer, 2008.

[34] R. G. Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968.

[35] B. Barbour, N. Brunel, V. Hakim, J. Nadal, *What can we learn from synaptic weight distributions?*, Trends Neurosci 30(12): 622-629, 2007.

[36] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Edited by A. Jeffrey and D. Zwillinger. Academic Press, New York, 7th edition, 8.447, p. 961, 2007.

[37] V. P. Singh and K. Singh, *Derivation of the gamma distribution by using the principle of principle of maximum entropy (POME)*, Water Resour. Bull. 21 (6): 941-952, 1985.

[38] A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford Studies in Probability, No. 2. Oxford University Press, 1992.

[39] T. Berger, *Rate Distortion Theory*. Prentice-Hall, Englewood Cliffs, 1971, Section 6.4.

[40] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Izd. Akad. Nauk. SSSR, Moscow, 1964. (Translated by A. Feinstein, Holden-Day, San Francisco.)

[41] C. E. Shannon, *A Mathematical Theory of Communication. Bell Sys. Tech. J., Vol. 27, 379-423, 623-656*, July and October, 1948.

[42] G. L. Gerstein, B. Mandelbrot, *Random walk models for the spike activity of a single neuron*, Biophys J. 4:41-68, 1964.

[43] O. E. Barndorff-Nielsen, P. Blaesild and C. Halgreen, *First hitting time models for the generalized inverse Gaussian distribution*, Stochastic Process and Their Applications, vol. 7, pp. 49-54, 1978.

[44] O. E. Barndorff-Nielsen and P. Blaesild, *Hyperbolic distributions and ramifications: contributions to the theory and application*, Statistical Distributions in Scientific Work. (C. Tallie et al., Eds.) 4 Reidel, Dordrecht, pp. 19-44, 1981.

[45] E. L. Brown, R. Barbieri, U. T. Eden and L. M. Frank, *Likelihood Methods for Neural Spike Train Data Analysis*. Chapter 9 of Computational Neuroscience: a Comprehensive Approach, J. Feng, ed, CRC Press, 2004.

[46] S. Iyengar and Q. Liao, *Modeling neural activity using the generalized inverse Gaussian distribution*, Biol. Cyber., vol. 77, pp. 289-295, 1997.

[47] E. Schrödinger, *Zur Theorie der Fall - und Steigversuche an Teilchenn mit Brownscher Bewegung*, Physikalische Zeitschrift, vol. 16, pp. 289-295. 1915.

[48] M. Abramowitz and I. A. Segun. *Handbook of Mathematical Functions*, Dover Publications, New York. 1965.

[49] T. Berger and W. B Levy, *Energy-Efficient Encoding of Information*, NSF Proposal 1162449, submitted September, 2011.

[50] F. Dyson, *A meeting with Enrico Fermi*, Nature, 427(6972): 297, 2004.

[51] J. Mayer, K. Khairy and J. Howard, *Drawing an elephant with four complex parameters*, Am J Phys, 78:648-649, 2010.

[52] H. B. Barlow, *Possible principles underlying the transformation of sensory messages*, In Sensory Communication (ed. Rosenblith, W. A.), 217-234, MIT, Cambridge, 1961.

[53] L. C. Sincich, J. C. Horton, T. O. Sharpee, *Preserving information in neural transmission*, J. Neurosci. 29, 6207-6216,2009.

[54] D. L. Rathbun, D. K. Warland, W. M. Usrey, *Spike timing and information transmission at retinogeniculate synapses*, J. Neurosci. 30: 13558-13566,2010.

[55] X. Wang, J. A. Hirsch, F. T. Sommer, *Recoding of sensory information across the retinothalamic synapse*, J. Neurosci. 30: 13567-13577,2010.

[56] J. A. Hirsch, J. M. Alonso, R. C. Reid, L. M. Martinez, *Synaptic integration in striate cortical simple cells*, J. Neurosci. 18, 95179528, 1998.

[57] L. M. Martinez, Q. Wang, R.C. Reid, C. Pillai, J.M. Alonso, F. T. Sommer, J.A. Hirsch, *Receptive field structure varies with layer in the primary visual cortex*. Nature Neurosci. 8, 372379, 2005.

[58] X. Wang, Y. Wei, V. Vaingankar, Q. Wang, K. Koepsell, F. T. Sommer, J.A. Hirsch, *Feedforward excitation and inhibition evoke dual modes of firing in the cat's visual thalamus during naturalistic viewing*, Neuron 55, 465 478, 2007.

[59] F. Briggs, W. M. Usrey, *Corticogeniculate feedback and visual processing in the primate*, J. Physiol. 589, 33-40, 2011.

[60] F. Briggs, W. M. Usrey, *Parallel processing in the corticogeniculate pathway of the macaque monkey*, Neuron 62, 135-146, 2009.