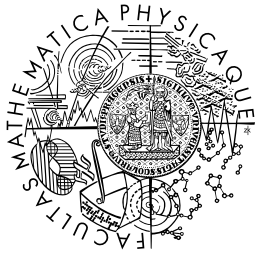

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Tomáš Bárta

Information - theoretical properties of chosen stochastic neuron models

Institute of Physiology of the Czech Academy of Sciences

Supervisor of the master thesis:	Mgr. Lubomír Košťál, PhD.
Study programme:	Mathematical and Computer Modeling in Physics
Study branch:	Physics

Prague 2018

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

Title: Information - theoretical properties of chosen stochastic neuron models

Author: Tomáš Bárta

Department of Computational Neuroscience: Institute of Physiology of the Czech Academy of Sciences

Supervisor: Mgr. Lubomír Kostál, PhD., Department of Computational Neuroscience

Abstract: According to the classical efficient-coding hypothesis, biological neurons are naturally adapted to transmit and process information about the stimulus in an optimal way. Shannon's information theory provides methods to compute the fundamental limits on maximal information transfer by a general system. Understanding how these limits differ between different classes of neurons may help us to better understand how sensory and other information is processed in the brain. In this work we provide a brief review of information theory and its use in computational neuroscience. We use mathematical models of neuronal cells with stochastic input that realistically reproduce different activity patterns observed in real cortical neurons. By employing the neuronal input-output properties we calculate several key information-theoretic characteristics, including the information capacity. In order to determine the information capacity we propose an iterative extension of the Blahut-Arimoto algorithm that generalizes to continuous input channels subjected to constraints. Finally, we compare the information optimality conditions among different models and parameter sets.

Keywords: information capacity, neuronal model, frequency coding

First and foremost I would like to thank my thesis supervisor Lubomír Košťál who led me very actively and patiently guided me through the fields of computational neuroscience and information theory. I also owe big thanks my parents who supported me throughout my studies.

Contents

Introduction

The presented work belongs to a relatively young field called *computational neuroscience*. As defined in the classical paper [1]:

> The ultimate goal of computational neuroscience is to explain how electrical and chemical signals are used in the brain to represent and process information.

The methodological tools employed by this field include (but do not restrict to) biophysics, theory stochastic processes and information theory. Different kinds of mathematical and physical models of the nervous system and its individual parts are constructed in order to better describe, quantify and interpret the processes in the brain.

The models used to investigate and describe nervous system are divided into three classes:

- **Descriptive (or formal / statistical) models** - the main goal of descriptive models is to reproduce experimental data. Such models then allow us to analyze the behavior of the system without considering the exact underlying biophysics. An example can be modeling the intervals between spikes (electrical signals neurons use to communicate) as being drawn from a parametric distribution (e.g. gamma or Weibull).

- **Mechanistic (biophysical) models** try to describe how the nervous system works by considering, e.g., its anatomy and to derive its properties from the "first principles". A classical example is the Hodgkin's and Huxley's model of a nervous cell, for which the authors have been awarded the 1963 Nobel prize in Physiology or Medicine.

- **Interpretive (functional) models** help us understand why the nervous cells behave the way we observe by assuming that the main purpose of the cells is to convey and process information. Neurons in the brain are stimulated by input either from other neurons connected to them or from the external environment and the neuron's response should reflect the input [2]. This process of "neural coding" is usually interpreted by the means of Shannon's information theory [3] and statistical estimation theory [4].

The presented work focuses on the level of individual cells and essentially involves all three of the above mentioned modeling approaches. First we use the recently proposed Multi-Timescale Adaptive Threshold (MAT) neuronal model [5], which includes both the statistical (formal) and biophysical components, to describe the stimulus-response relationship. Despite its simplicity, the MAT model was shown to reproduce spike trains of a wide range of neurons very well [5], [6]. Second, we calculate the ultimate limits on reliable information transmission by these neurons.

Shannon developed a theory for transmission of information in electrical systems and this theory has been successfully applied to a wide range of neuroscientific problems, as reviewed for example in [7]. In the framework of the Shannon

information theory, a neuron is treated as an information channel. Information theory allows us to quantify the information a channel transmits based on the statistical properties of its input and output and sets upper bound on the transmission rate, which cannot be surpassed. It is believed that the neural system is able to transmit information very near this theoretical bound.

We compare these limits between different classes of neurons represented by different parameters of the MAT model. The novelty brought by this work is the comparison of information-theoretical properties for the different neuronal activity regimes and their metabolic efficiency.

1. Neural cells and their mathematical models

1.1 Structure of a neuron and basic electrical properties

The brain is composed of cells called neurons. These cells are often interconnected. They process and transmit information (e.g. from our eyes) and their activity leads to an appropriate reaction (e.g. muscle movement). All tissue cells (not only neurons) have a bilipid membrane impermeable neither to water nor electric ions. The membrane contains ion pumps and ion channels that maintain a difference between electric potentials outside and inside the cell. However, the membrane of a neuron is electrically excitable, meaning that it's properties (i.e. the conductance) are strongly dependent on the potential difference. This key property enables neurons to communicate by propagating electrical signals in the form of so called *action potentials*, also called *spikes*.

In the figure 1.1 a schematic drawing of a neuron is presented. For the description of signal transfer in neurons, dendrites and axon are the most important parts to understand.

- **Dendrites** are the parts of a neuron that receive signal from other neurons (some neurons, e.g. sensory, are equipped with specialized receptors that ultimately transform the external signal, e.g. the chemical concentration as in the case of olfactory sensory neurons, into a change in the membrane potential). Dendrites often branch out in a tree like fashion.

- **Axon** is the part of a neuron that transfers signal to other neurons. At the axon terminals the neuron connects to dendrites of other neurons with **synapses** (i.e. contact points).

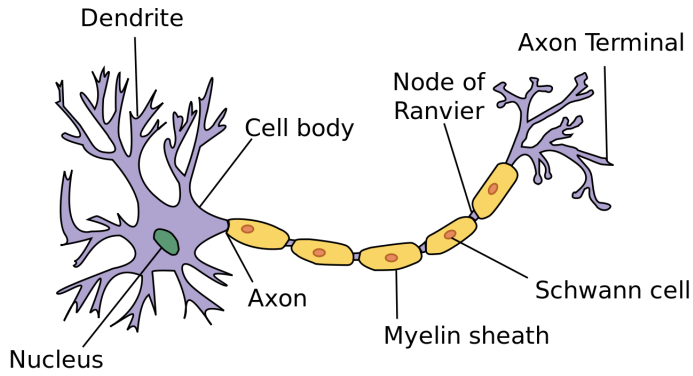

Figure 1.1: Simple schematic of a neuron, credits: [8]

The two neurons connected by a synapse are call the pre-synaptic and the post-synaptic neuron - the axon of the pre-synaptic neuron is connected to the dendrite of the post-synaptic neuron. A chemical synapse (which is by far the

most common type of synapse) is activated by a specific molecule (a neurotransmitter), which is released by the pre-synaptic neuron. The activation results into a change of the membrane potential of the neuron - the mechanism will be discussed in more detail in the section 1.4.

When a synapse is activated, so called post-synaptic potential then propagates through the dendrites of the receiving neuron and when certain conditions are met, the neuron fires a spike. The spike is then conducted through its axon, possibly causing a post-synaptic action potential in other neurons, if they're connected through a synapse.

1.2 Neuron as an electrical circuit

For the purpose of this thesis we neglect the detailed neuronal anatomy and ignore the uneven distribution of the membrane potential. Models using this simplification are called single-compartment or single-point models. Even such simplified models, however, were shown to reproduce a rich variety of realistic neuronal responses remarkably well [9], [10]. If not stated otherwise, the information in this section are drawn from [2].

1.2.1 Equilibrium and reversal potentials

There is typically an excess of negative charge inside the neural cell, relative to its surroundings. This imbalance is maintained by ionic pumps that expand energy (in the form of ATP molecules) to actively move negatively charged ions into the cell. Diffusion drives ions from the place with higher concentration to the place with lower concentration (Na^+, Ca^+ are driven into neuron, K^+ out of the neuron).

There are many different types of ion channels present in the membrane, some are selective only to a single type of ions (e.g. Na^+), some can allow more types to pass through. Diffusion will drive ions either into the cell or out of the cell, depending on the concentration gradient. To be able to leave or enter the cell, the ion has to have a sufficient energy to overcome the energy barrier produced by the membrane potential.

For each type of channel we may define the equilibrium potential, called reversal potential, at which the rate of ions entering the neuron is the same, as the rate of them leaving. The potential difference between the inside and the outside of a neuron when at rest (i.e. no signal is received from outside) is called the resting potential and is the result of a combination of many different channel types in the neuron's membrane. A neuron's resting potential is about $V_{rest} = -70\,mV$.

Ion channels effectively act as a conductance (or resistance), that tend to move the membrane potential towards the resting potential. When the potential difference V is lower than the channel's reversal potential E, positive charge will enter the neuron through this channel, thus increasing the potential difference (depolarizing the cell) and vice versa.

1.2.2 Membrane capacitance and resistance

The potential gradient described above causes the negative charge inside the cell and positive charge outside of the cell to accumulate on the membrane. Thus, the membrane acts as a capacitance. This membrane capacitance is denoted as C_m and can be expressed as

$$C_{\mathrm{m}} = \frac{Q}{V}, \tag{1.1}$$

where Q is the excess charge and V is the potential difference (voltage). This equation plays a key role in all single-compartmental mathematical models of neuronal activity. The total capacitance is proportional to the area A of the neuron, which allows us to define the specific capacitance c_{m}:

$$c_{\mathrm{m}} = \frac{C_{\mathrm{m}}}{A}. \tag{1.2}$$

Remarkably, the specific capacitance is approximately the same for all neurons, $c_{\mathrm{m}} \approx 10\,\mathrm{nF\,mm^{-2}}$, the membrane surface area ranges from about $0.01\,\mathrm{mm^2}$ to $0.1\,\mathrm{mm^2}$.

As mentioned above, the ion channels act as conductances - therefore we can assign a resistance R_{m} to the membrane. Then following the Ohm's law, if we want to keep the potential difference at a value different from the cell's resting potential, we have to inject a current I_{e}, thus changing the voltage by

$$\Delta V = I_{\mathrm{e}} R_{\mathrm{m}}, \tag{1.3}$$

The membrane resistance is inversely proportional to the membrane area (greater area leads to more current-conducting ion channels), allowing us to define the specific membrane resistance r_m, where

$$R_{\mathrm{m}} = \frac{r_{\mathrm{m}}}{A}. \tag{1.4}$$

The product of the membrane resistance and the membrane capacitance is called the membrane time constant:

$$\tau_{\mathrm{m}} = R_{\mathrm{m}} C_{\mathrm{m}} = r_{\mathrm{m}} c_{\mathrm{m}} \tag{1.5}$$

The membrane time constant determines the time scale on which the voltage changes when a current is injected (just like in an RC circuit). The values of time constants in neurons typically range from 10 ms to 100 ms.

1.2.3 Total membrane current

The current flowing through any channel type obeys the Ohm's law:

$$I_i = \frac{V - E_i}{R_i} \tag{1.6}$$

$$i_i = g_i(V,t)(V - E_i) \tag{1.7}$$

where the subscript i denotes channel type, E_i the reversal potential for this channel type, R_i is the resistance and I_i the total current flowing through these

channels, i_i and g_i are the specific current ($i_i = I_i/A$) and the specific conductance - generally dependent both on time and on the potential difference.

The total (specific) current i_m flowing through all different channels can be computed as a sum over all channel types:

$$i_\mathrm{m} = \sum_i g_i(V,t)(V - E_i). \tag{1.8}$$

The remaining sources of electrical current, including the ion pumps, are usually grouped together and the resulting current is called the leakage current, denoted as i_L. The leakage current is modeled analogously to the ion channel currents, with its own specific conductance $\bar{g}_\mathrm{L}$ (the bar indicates that the value is independent of voltage) and reversal potential E_L:

$$i_\mathrm{L} = \bar{g}_\mathrm{L}(V - E_\mathrm{L}). \tag{1.9}$$

From equation (1.1) follows for a constant specific membrane capacitance c_m

$$\begin{aligned} c_\mathrm{m}\tfrac{\mathrm{d}V}{\mathrm{d}t} = \tfrac{\mathrm{d}Q}{\mathrm{d}t} = i_\mathrm{m} + i_\mathrm{L} = i_\mathrm{tot} \\ V(t=0) = V_\mathrm{rest} \end{aligned} \tag{1.10}$$

This equation is common for all single-compartment models.

1.2.4 Hodgkin-Huxley model

Hodgkin-Huxley model describes the neuron as an electrical circuit with two different types of electrically active ion channels - namely K^+ channels and Na^+ channels. The channels have specific structures and the Hodgkin-Huxley model takes advantage of this structure to describe the voltage and time dependence of their conductances.

The significance of this model lies in its ability to reproduce the full time course of an action potential - a sudden rapid depolarization of the neuron's membrane propagating through the axon towards other neurons.

In the figure 1.2 the time course of membrane voltage in the Hodgkin-Huxley model is shown, when an external current is injected. When the current is strong enough a spike is fired. The shape of the spike is always the same.

1.3 Integrate-and-Fire models

Only the action potential is capable of triggering a synapse and thereby exciting another neuron. Moreover, observing that the time course of an action potential for a given neuron is always the same and that its duration is much shorter than the intervals between successive action potentials, leads to the conclusion, that for the information transfer only the times of occurrence of individual action potentials (or spikes) matter. A neuron's activity can then be described by a so called spike train - i.e. a list of times when spikes occur in a given time interval.

When only the spike train is of interest, simplified models can be used instead of detailed biophysical models like the Hodgkin-Huxley model. Important class of such models are the integrate-and-fire models. These models avoid describing

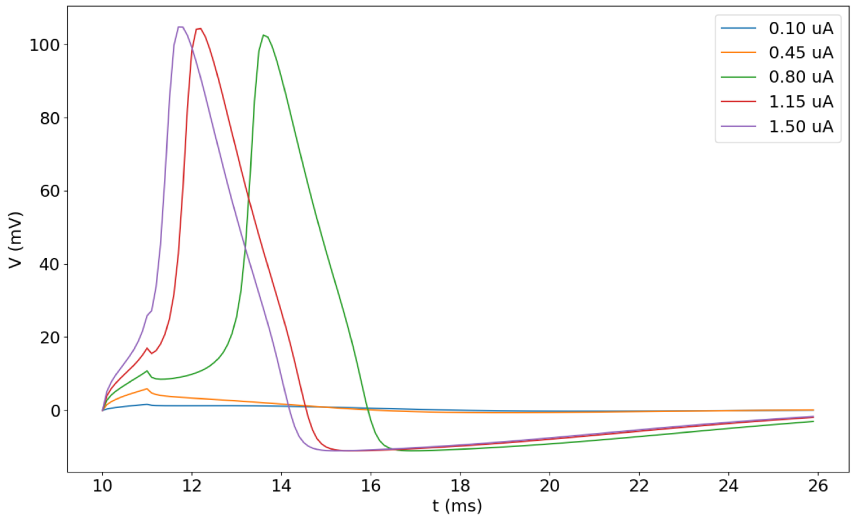

Figure 1.2: Membrane voltage behavior when step currents of different magnitudes are introduced to the Hodgkin-Huxley model for 1 ms. Strong enough input current results in an action potential. It can be seen that the shape of the action potential is in all cases practically the same. The model was simulated using the Brian 2 package and the example code provided in its documentation [11]

.

the action potential when a spike is fired, instead they set a threshold voltage θ and only behavior under this voltage is modeled. When the membrane potential reaches the threshold θ a spike is fired and the membrane potential is set to the value of the resting potential V_{rest}.

Next, some basic models of an integrate-and-fire neuron describing the behavior of the neuron in the sub-threshold domain will be described.

1.3.1 Perfect Integrate-and-Fire neuron

In the simplest case the membrane is considered to be a perfect insulator, leading to no exchange of charge between the neuron and its surroundings. Since we are talking about single-compartment models, equation (1.10) holds with $i_{\text{tot}} = 0$. When we take an externally injected current I_{e}, the equation describing the sub-threshold behavior of V takes on the form

$$\begin{aligned} c_m \frac{\mathrm{d}V}{\mathrm{d}t} &= \frac{I_{\text{e}}}{A} \\ V(t=0) &= V_{\text{rest}} \end{aligned} \tag{1.11}$$

Thus, all the charge from the external current stays on the capacitor (membrane), until the voltage on the capacitor reaches the threshold value θ.

Solution of this equation is obtained by integration:

$$V(t) = V(t_0) + \frac{1}{C_{\text{m}}} \int_{t_0}^{t} I_{\text{e}} \, \mathrm{d}t \tag{1.12}$$

1.3.2 Leaky Integrate-and-Fire neuron

When only the leakage conductance $\bar{g}_{\mathrm{L}}$ is taken into account, only the leakage current $i_{\mathrm{L}} = \bar{g}_{\mathrm{L}}(V - E_{\mathrm{L}})$ remains and equation (1.10) becomes

$$c_m \frac{\mathrm{d}V}{\mathrm{d}t} = -\bar{g}_{\mathrm{L}}(V - E_{\mathrm{L}}) + \frac{I_{\mathrm{e}}}{A}, \tag{1.13}$$

which can be also expressed as

$$\tau_{\mathrm{m}} \frac{\mathrm{d}V}{\mathrm{d}t} = E_{\mathrm{L}} - V + R_{\mathrm{m}} I_{\mathrm{e}}. \tag{1.14}$$

According to convention the leakage current i_{L} is said to be positive, if it flows inside the neuron, whereas the membrane current is positive if it flows from inside the cell to the outside.

In the sub-threshold domain the general solution of this equation is

$$V(t) = E_{\mathrm{L}} + \frac{R}{\tau_{\mathrm{m}}} \exp\left(-\frac{t}{\tau_{\mathrm{m}}}\right) * I(t), \tag{1.15}$$

where $*$ denotes the convolution.

The model described by (1.13) is called the Leaky Integrate-and-Fire (LIF) and was introduced by Stein in 1965 [12]. The most notable difference from the non-leaky (perfect) integrate-and-fire neuron is that when no external current is present, the potential decays towards its resting value.

The LIF model can be further extended in order to account for other possible properties which it fails to replicate in its simple form.

Adaptive Integrate-and-Fire neuron

Typically, the firing rate of neurons exposed to a steady stimulus slowly declines as the neuron adapts to the conditions. One possible approach to modeling this phenomena is the introduction of an additional K^+ conductance g_{sra} [13] (with the reversal potential of about $-77\,\mathrm{mV}$).

When a spike is fired, this conductance increases by Δg_{sra}, thus hyperpolarizing the neuron and inhibiting the firing process. Otherwise, this conductance decays exponentially with a time constant τ_{sra}.

In the figure 1.3 responses of leaky integrate-and-fire units without and with spike rate adaptation to a step current are compared.

Refractory period

An experimental evidence shows that a neuron is very unlikely to fire a spike right after one has just occurred. This is already implicitly included in the Hodgkin-Huxley model, but needs to be added to the integrate-and-fire models separately. In the integrate-and-fire models the refractory period can be achieved for example by prohibiting any voltage change (or any new spike to be fired) in the neuron shortly after a spike is fired or by introducing a mechanism similar to the spike rate adaptation, but with a much shorter time constant. A typical value of the refractory period is 2 ms

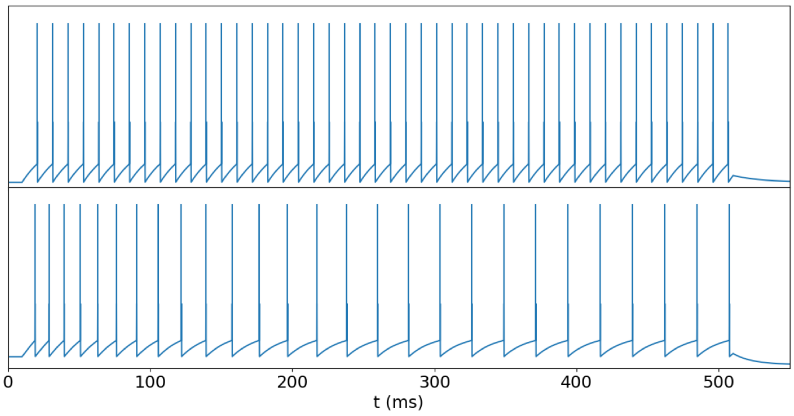

Figure 1.3: Reaction of the leaky integrate-and-fire models to a step current $0.6\,\mu$A with a duration of 500 ms without (upper) and with (lower) spike rate adaptation (s.r.a.) mechanism modeled as a K^+ conductance. The y axis shows the voltage, vertical line is shown when a spike is generated. The interspike intervals of the model with s.r.a. begin as a bit shorter, but after first few spikes the unit fires regularly, whereas the unit without s.r.a. fires regularly from the beginning. Parameters common for both models are: $C_\mathrm{m} = 0.5\,\mu\mathrm{F}$, $V_\mathrm{th} = -45\,\mathrm{mV}$. For model without s.r.a., the membrane resistance was set to $R_\mathrm{m} = 40\,\mathrm{M\Omega}$, for model with s.r.a. $R_\mathrm{m} = 100\,\mathrm{M\Omega}$, $g_\mathrm{sra} = 0.003\,\mu\mathrm{S}$, $\tau_\mathrm{sra} = 100\,\mathrm{ms}$, $V_\mathrm{sra} = -80\,\mathrm{mV}$.

1.4 Synapse modeling

Synapses were already briefly described in 1.1. They are the contact points between neurons and the prevalent type - the chemical synapse - is activated by neurotransmitters. When a pre-synaptic neuron fires a spike, the action potential propagates through its axon and when it reaches one of the axon's terminal, neurotransmitters might be released and bind to the receptors on the post-synaptic neuron, thus opening chemically sensitive ion channels and due to the ion flow, an electrical current can be observed.

Depending on the synapse reversal potential, E_syn (i.e. which type of channels is opened), the synapse can be either excitatory or inhibitory. Reversal potential of an inhibitory synapse is lower than the post-synaptic neuron's resting potential and the post-synaptic neuron is hyperpolarized when the synapse is activated. Reversal potential of an excitatory synapse is higher, which causes a depolarization of the neuron.

Mathematically, opening of a synapse's ion channels can be expressed as an increase in the synapse's conductance g_s, which can be in most cases expressed as

$$g_\mathrm{s} = P_\mathrm{s}\bar{g}_\mathrm{s}, \tag{1.16}$$

where P_s denotes the probability of a single channel being open (therefore also the fraction of all the open channel), $\bar{g}_\mathrm{s}$ is the conductance if all the synapse's channels were open.

The dynamics of P_s is ruled by following differential equation:

$$\begin{aligned} \tfrac{\mathrm{d}P_\mathrm{s}}{\mathrm{d}t} &= \alpha_s(1-P_\mathrm{s}) - \beta_\mathrm{s}P_\mathrm{s}, \\ P(t=t_\mathrm{s}) &= 0 \end{aligned} \tag{1.17}$$

where α_s represents the rate at which channels go from the open to the closed state (*opening rate*), β_s the *closing rate* and t_s denotes the time of the beginning of the process.

The coefficient α_s depends heavily on the amount of available neurotransmitters nearby. The neurotransmitters can be present for a very brief amount of time until they diffuse in the surroundings. This results in an initial rise in P_s and when the neurotransmitters are no longer present, P_s starts to drop exponentially towards 0.

If we consider α_s to be a step function ($\alpha_\mathrm{s} = \bar{\alpha}_\mathrm{s}$ for $t \in [0,T]$ and 0 otherwise, where $\bar{\alpha}$ is a positive constant) and β_s is constant, we obtain the following solution:

$$P_\mathrm{s}(t) = \begin{cases} \frac{\alpha_\mathrm{s}}{\alpha_\mathrm{s}+\beta_\mathrm{s}}\left(1-e^{-(\alpha_\mathrm{s}+\beta_\mathrm{s})t}\right) & t \in [0,T] \\ P_\mathrm{s}(T)e^{-\beta_\mathrm{s}(t-T)} & t > T, \end{cases} \tag{1.18}$$

where $P_\mathrm{s}(T) = \frac{\bar{\alpha}_\mathrm{s}}{\bar{\alpha}_\mathrm{s}+\beta_\mathrm{s}}\left(1-e^{-(\bar{\alpha}_\mathrm{s}+\beta_\mathrm{s})T}\right)$ can be thought of as P_max - the value at which P_s peaks.

For some types of receptors (e.g. the AMPA[1] receptor), the initial increase in P_s is so rapid, that this phase can be completely neglected and the function takes on a simple form:

$$P_\mathrm{s}(t) = P_\mathrm{max}e^{-\beta_\mathrm{s}t}. \tag{1.19}$$

When all synapses are considered to be independent (as is the usual practice), we can compute the total conductance as the sum of all the individual conductances. For identical AMPA synapses the total conductance can be expressed as

$$g_\mathrm{AMPA}(t) = \bar{g}_\mathrm{AMPA}\sum_{k=1}^{N} P_\mathrm{max}e^{-(t-t_k)/\tau_\mathrm{AMPA}}, \tag{1.20}$$

where t_k are times of occurrence of individual synapses and the resulting current is then

$$I_\mathrm{AMPA}(t) = g_\mathrm{AMPA}(t)(V - E_\mathrm{syn}). \tag{1.21}$$

[1]Named after an artificial neurotransmitter, the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid, or AMPA

2. Multi-timescale adaptive threshold model

The multi-timescale adaptive threshold (MAT) model was proposed relatively recently in [5] as a refinement of the classical LIF model. The MAT model is relatively simple, yet it has been shown that it is able to predict spike times of many different types of neurons.

The MAT model is essentially the Leaky-Integrate-and-Fire model in the sense that the voltage is governed by the equation

$$\tau_{\mathrm{m}}\frac{\mathrm{d}V}{\mathrm{d}t} = -(V - E_{\mathrm{L}}) + RI(t), \tag{2.1}$$

i.e. the same equation as (1.13) but with time constant $\tau_{\mathrm{m}} = C_{\mathrm{m}}R_{\mathrm{m}}$. The difference from the LIF model is that the voltage doesn't change when a threshold is reached, but instead dynamics for the threshold are introduced.

In the MAT model, the threshold $\theta(t)$ is composed $L+1$ summed components:

$$\theta(t) = \omega + \sum_{j=1}^{L}\theta_j(t), \tag{2.2}$$

where ω is the constant *resting threshold value* and the components $\theta_j(t)$ decay exponentially towards 0 with time scales τ_j. Each of these components is increased by α_j each time a spike occurs.

The dynamics of the voltage threshold $\theta(t)$ are then described by

$$\theta(t) = \sum_{k} H(t - t_k) + \omega \tag{2.3}$$

$$H(t) = \sum_{j=1}^{L}\alpha_j \exp(-t/\tau_j), \tag{2.4}$$

where k iterates through all the previous spikes and t_k is the kth spike's time. Important detail is the use of the absolute refractory period 2 ms - after a spike is generated, another spike cannot occur for the next 2 ms. This prevents the model from a so called singular bursting, i.e. firing spike continuously when the membrane voltage exceeds the threshold for a longer period of time. Figure 2.1 depicts how the dynamics of the threshold determine when action potential occurs.

A model with L different time scales is referred to as MAT(L).

Kobayashi et al. illustrate the predictive ability of the MAT model on electrophysiological recordings from a wistar rats' cortex neurons. A current of a stochastic nature was injected into the neurons and the response was recorded.

The recordings were used to fit multiple versions of the MAT model, but the MAT(2) model was selected as the best general model with parameters common for all experiments $\tau_{\mathrm{m}} = 5\,\mathrm{ms}$, $R = 50\,\mathrm{M\Omega}$ and $\tau_1 = 10\,\mathrm{ms}$, $\tau_2 = 200\,\mathrm{ms}$. The time scales α_1 and α_2 will be also referred to as the fast and slow component of the model. In accordance with the original work this model will be further denoted

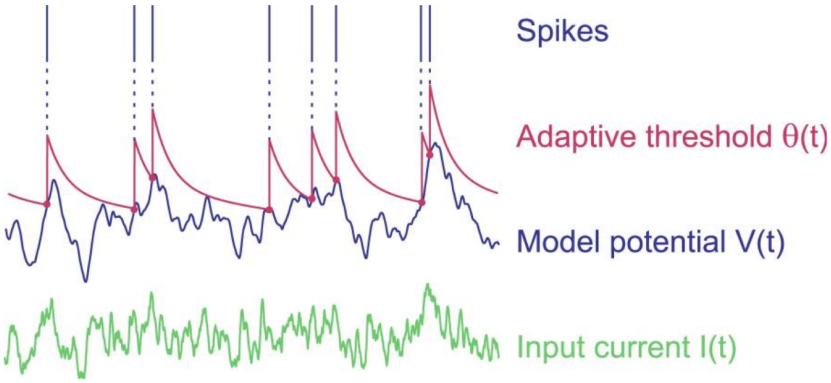

Figure 2.1: The model is stimulated by an input current. The change in voltage is evaluated by (2.1), the adaptive threshold follows the equations (2.3), (2.4). When the voltage reaches the value of the dynamic threshold a spike is fired and the threshold value jumps up. Credit: [14]

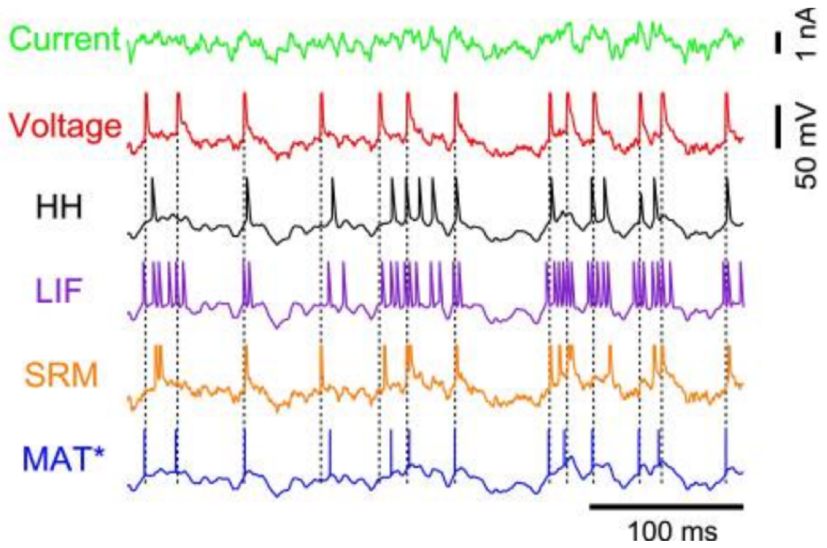

Figure 2.2: Comparison of ability of different models (Hodgkin-Huxley, LIF, Spike Response Model (SRM) and MAT) to reproduce times of action potentials, given the input current, to experimental recording. Credit: [5]

as MAT*. The parameters α_1, α_2 and ω were treated as free parameters and were fitted individually for each neuron. Figure 2.2 shows how the MAT* model predicts when a neuron is stimulated by a given input current and compares the prediction with reality next to other common neuron models.

It is obvious from the figure that the MAT* model performed the best. In the article [5] the authors continue to evaluate the predictive performance quantitatively in multiple different scenarios and the MAT was generally the best performing one.

3. The problem of neural coding

It is generally accepted that neurons transmit information by producing spikes (action potentials) [2], since spikes are the only mechanism allowing neurons to communicate with each other as was explained in chapter 1. For this reason, when only the transmitted information is of our interest, effective models like the integrate-and-fire models or MAT are sufficient. Even though they don't thoroughly follow the biophysics of a real neuron, they can predict the spike times very well [5].

When neurons are subjected to a stimulus, they pass on the information about the stimulus in the form of a spike train. It is still unknown how the information should be "read out" from the spike trains. However, some hypotheses were proposed. These can be categorized into two groups:

1. **Rate coding** - the information about stimulus is conveyed by the estimated firing rate (i.e. essentially the number of fired spikes) of a single neuron or a group of neurons

2. **Temporal coding** - the spike code encodes the information in the exact times of spike firing (e.g. in the time it takes for a neuron to fire the first spike after being introduced to a new stimulus)

3.1 Rate coding

The simplest definition of the mean firing rate of a neuron is the number of spikes fired by a neuron averaged over some time window (called the *coding time window*), specified by the experimenter (e.g. 100 ms or 500 ms).

Experiments proved that in some cases stronger stimulus indeed results in more spikes, thus allowing the experimenter to infer the stimulus strength from the number of spikes observed [15], [16].

In reality, however, the reaction times can be much faster than the above mentioned lengths of time windows would suggest. A classical example is a behavioral experiment showing that a fly can react to a new stimulus (by changing direction) within 30 ms to 40 ms [17]. This would suggest that even though the spike mean firing rate is related to the stimulus strength, the real decoding mechanism is different.

3.2 Temporal coding

In rate code, as mentioned above, everything in the spike train except for the number of spikes is considered redundant and is thrown away. In temporal coding, however, the exact timing of spikes is considered to carry information.

3.2.1 Latency coding

An example of coding strategy based on exact spike timing is latency coding (also called time-to-first-spike coding), where the transmitted information lies in the

time it takes for the neuron to fire a first spike when a new stimulus is presented. If neurons were using this scheme, all the information would be conveyed in the timing of the first spike after onset of the new stimulus.

The timing of the first spike has to be measured relative to some reference signal. A realistic situation in which such a coding strategy could be employed is when an eye is scanning a picture. The eye's focus shifts quickly between points on the picture in so called saccades. If each shift of focus happened after spike had been fired, the timing of this spike could then be used as a reference signal for the next spike.

3.2.2 Population spike coding

When we want to consider the timing of spikes to convey information, we always need to have a reference signal. In the example with eyes looking at a picture, the previous spike of the considered neuron was thought of as the reference signal. But we can also reference the timing to the behavior of other neurons. Possibilities include:

a. Spikes from other neurons - we can use a spike from neuron 1 as a reference signal for neuron 2

b. Oscillations - periodic behavior of groups of neurons is quite common in the brain, therefore it can be used by other neurons as a kind of clock

3.2.3 Reverse correlation

Reverse correlation is a technique which can be used to reconstruct the stimulus evoking the investigated spike train and it can help us understand, what makes the investigated neuron fire.

If we have a spike train, and we know what stimulus evoked this spike train, we can then extract for each spike the preceding stimulus (e.g. stimulus in a time window ranging 100 ms before the spike to the time of the spike) and average these recordings. Thus, we obtain a typical time course (we denote the time dependent function as $\kappa(t - t^{(f)})$, where $t^{(f)}$ is the time when the spike is fired) of the stimulus evoking a spike.

When given a new spike train from the same neuron (described by a set of firing times $\{t_i\}_{i=1}^{N}$), where the stimulus is unknown, we can estimate the complete time course of the stimulus as a linear combination of the functions obtained by the reverse correlation:

$$s_{\text{est}}(t) = \sum_{i=1}^{N} \kappa(t - t_i). \tag{3.1}$$

It has been shown that in certain cases this reconstruction technique can give very good estimates of the stimulus' time course [18].

4. Information theory in neuroscience

One of the major functions of the nerve cells, e.g. in the peripheral and sensory pathways, is the transmission of information about the environment to the central nervous system. In the previous chapter we discussed the classical hypotheses, i.e. neural coding paradigms describing how information can be represented by a neuron's spike train.

In this chapter we treat the neuron as an information channel, whose role it is to convey the received information. The rate coding hypothesis states, that a neuron can transmit information about its input by altering the rate at which it fires spikes. In an ideal case we would be able to tell by the response what is the input of the neuron. In reality, however, we usually can't tell exactly from the response what is the input, because one input does not always produce the same response over repeated trials [2]. Such neuron corresponds to the notion of a so called noisy information channel.

It was shown by Shannon that such channels are still capable of reliably conveying information and Shannon's information theory became the standard way to investigate the properties of neurons as information channels.

The use of Shannon's information theory is motivated by a so called Efficient coding hypothesis [19], which states that neurons are developed by evolutionary processes to convey information as efficiently as possible. This hypothesis predicts that the statistical characteristics of input and output of individual neurons should satisfy those required by the information-theoretical optima, as supported by multiple experiments [20], [21].

4.1 Framework of Shannon's information theory and its relation to nervous cells

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.
>
> —Claude Shannon, A Mathematical Theory of Communication

As depicted in the diagram in the figure 4.1, a general communication system consists of 3 parts (not counting the source and the destination) - an encoder,

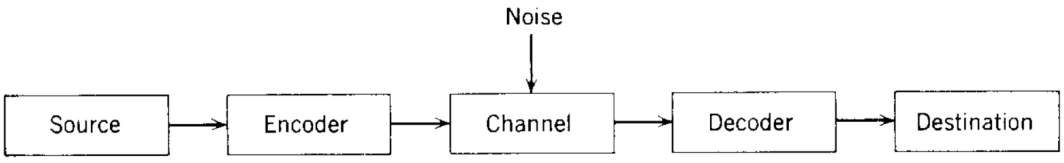

Figure 4.1: A general scheme of a point-to-point communication system. Image taken from [22]

a channel and a decoder. The output of the source can be anything we wish to convey, e.g. a voice waveform or a sensory input. The message is then transmitted by a channel and we want to decode as accurately as possible what the message was.

To use the information channel, it is sometimes convenient and sometimes necessary to encode the message first. In practice, the encoder is usually further separated into two parts - source encoder and channel encoder. The source encoder encodes the message into a sequence of 0s and 1s. The channel encoder then maps these sequences onto symbols that can be transmitted by the channel.

The decoder can be again split into two parts - a channel decoder and a source decoder. The role of the channel decoder is to infer from the output of the channel what message entered the channel and then tries to reproduce the binary sequence. If the decoded message differs from the message generated by the source, we say that a *decoding error* has occurred. The source decoder then transforms the binary sequence into a specified form to the destination.

A possible cause for the decoding error can be the presence of a noise in the channel. Ideally, the channel would be a bijective map - for each possible output symbol of the channel there would be a unique symbol on the input capable of producing it. This is, however, usually not the case and each input symbol can be mapped to any output symbol with assigned probabilities.

One of the main results of information theory is the *noisy channel coding theorem*, which states, that even channels subjected to a noise can transmit information while the probability of a decoding error is arbitrarily low. Moreover, it provides us a quantity called the *information capacity* setting the maximal possible transmission rate.

In this work, we want to evaluate the information capacity of certain neuron models, therefore, since we are going to treat a neuron as an information channel, we are interested mostly in the properties of information channels. However, in order to better understand the whole picture we will also discuss the other parts of the communication system.

Also note, that the separation of an encoder into a source encoder and a channel encoder is done solely for practical reasons. It allows to completely separate the source from the channel and to get an intuitive feel for the notion of information.

4.1.1 Source models and source coding

In information theory the source is always modeled as a random process. In this work we will only consider discrete-time memoryless sources. The output of a discrete-time memoryless source is modeled as a sequence of independent observations from a set with a specified probability measure.

An example of a source can be any language. Of course, a message in any language isn't just a random sequence of letters, but it is a very useful simplification. In the case of English, the letters are chosen from the English alphabet of 28 letters with some additional symbols. Each letter occurs with a fixed frequency, given by the language. This can be taken advantage of when compressing messages in the given language. E.g. the Morse code uses a single dot for the most common letter in the English language - e, and a single dash for the second

most common letter - t [23]. This is quantified by the *Shannon entropy*.

Given a discrete random variable X on a set $\{x_1, ..., x_N\}$with a probability distribution p represented by a vector of probability assignments $(p_1, ..., p_N)$, the Shannon entropy $H(X)$ is defined as

$$H(X) = -k \sum_{j=1}^{N} p_j \ln p_j, \tag{4.1}$$

where k is an arbitrary constant, in information theory commonly chosen as $\log_2 e$ (we will also use this value). The Shannon entropy $H(X)$ is the minimal number of bits one has to use per one symbol on average to represent a sequence of i.i.d. observations of the random variable X.

As an example, consider a source which is able to send four different letters: *A*, *B*, *C*, *D*. The frequencies of those letters are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{8}$. Consider the two following possibilities to encode the letters:

	method 1	method 2
A	00	0
B	01	10
C	10	110
D	11	111

Using the first method, we always need two bits to encode a letter. Therefore the average number of bits required per symbol is 2. In the latter case the number of bits required to encode each letter differs and we need to take into account their frequencies. The average number of bits required to encode one letter is then

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{7}{4}. \tag{4.2}$$

I.e. on average $\frac{7}{4}$ of bits per symbol will be used instead of 2. Note that this is also the Shannon entropy of the corresponding probability distribution. It isn't possible to represent a message from such source while using less than $\frac{7}{4}$ bits per symbol on average. One could propose an apparently more efficient representation, e.g. *A*: 0, *B*: 1, *C*: 01, *D*: 11. However, using this representation, one can't locate the separations between letters. A sequence 011 can be either read as *ABB*, *AD* or *CB*.

The intuition behind entropy is that it measures the uncertainty of a random variable - it is the only function that satisfies three properties one would expect from a measure of uncertainty.

Theorem 1. *Let X, Y be discrete random variables with associated probability assignments $p = \{p_1, ..., p_N\}$ and $q = \{q_1, ..., q_M\}$. The Shannon entropy H of a discrete variable is uniquely defined by the following properties (up to a constant factor k):*

1. *$H(X)$ is continuous in p_i $\forall i \in 1, ..., N$*
2. *If p is uniform, i.e. $p_1 = p_2 = ... = p_N = \frac{1}{N}$, $H(X)$ is a monotonically increasing function of N. For a given N, $H(X)$ attains its maximum at this point.*

3. *Consider a discrete random variable Z_t with associated probability distribution $g = \{tp_1, ..., tp_N, (1-t)q_1, ..., (1-t)q_M\}$, $t \in [0,1]$. Then $H(Z) = tH(X) + (1-t)H(Y)$.*

The proof can be found in [3] or [24].

The source produces symbols at rate ν_s (in symbols per second). Therefore if a memoryless discrete-time source is described by a random variable X we can say that the source produces $\nu_s H(X)$ bits per second.

The Shannon entropy is not well defined for continuous random variables. This makes a good sense, since it is not possible to represent numbers from an interval of real numbers by a finite number of bits - such source produces infinite number of bits per second. The Shannon entropy can be generalized to continuous random variables by quantization (i.e. discretization). However, the detailed treatment is not the aim of this work.

4.1.2 Channel models

We are concerned in *reliable transmission* through the information channel. Reliable transmission means that by encoding the output of the source encoder well enough, we are able to keep the probability of the decoding error arbitrarily low. In the case of a noiseless channel, the reliability of the transmission is not an issue. However, a general information channel is subjected to noise.

Generally, a *discrete-time information channel* has a set of possible inputs $\mathcal{A}$ called the input alphabet and a set of possible outputs $\mathcal{B}$ called the output alphabet. The channel encoder presents the channel with a message $\mathbf{x} = (x_1,\ x_2,\ ...,\ x_N) \in \mathcal{A}^N$. The channel maps the input sequence to a sequence $\mathbf{y} = (y_1,\ y_2,\ ...,\ y_N) \in \mathcal{B}^N$ with probability $P_N(\mathbf{y}|\mathbf{x})$. The channel is called memoryless without feedback if each symbol in the output sequence depends only on the corresponding symbol in the input sequence, and we can therefore write

$$P_N(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} P(y_n|x_n), \tag{4.3}$$

where $P(y|x)$ is the probability of observing the output symbol $y \in \mathcal{B}$ when given an input symbol $x \in \mathcal{A}$. The conditional probability distribution $P(y|x)$ then provides a complete description of the discrete-time memoryless channel without feedback.

Until Shannon's result it was thought that such noisy channels couldn't be used for arbitrarily reliable communication. Shannon showed that communication through such channels is possible while keeping the probability of the decoding error as low as desired. However, we can't transmit more than C (in bits) per single channel use, where C is the information capacity of the channel.

Suppose the information channel can transmit symbols at frequency ν_c. If we want to transmit reliably a message from a source producing $\nu_s H(X)$ bits per second, we need a channel with an information capacity C satisfying:

$$C > \frac{\nu_s}{\nu_c} H(X). \tag{4.4}$$

4.1.3 Block code and the operational interpretation of the channel capacity

Block code is a coding scheme which is useful for understanding the notion of information capacity from the operational point of view. The idea is to process whole sequences of input values, $\mathbf{x} = (x_1,\ x_2,\ ...,\ x_N) \in \mathcal{A}^N$, rather than single values. There are altogether $|\mathcal{A}|^N$ possible sequences of length N. In block coding, a subset of M of those sequences $\mathcal{S} \subset \mathcal{A}^N$ is considered ($|\mathcal{S}| = M$). The rate R of information transmission in bits per second is then defined as

$$R = \nu_c \frac{\log_2 M}{N}. \tag{4.5}$$

The input sequence $\mathbf{x} \in \mathcal{S}$ is mapped by the channel to an output sequence $\mathbf{y} \in \mathcal{B}^N$. The role of the decoder is to infer which of the M possible sequences was on the input. The *maximum likelihood decoder* chooses such $\mathbf{x}^* \in \mathcal{S}$ that

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{S}} \mathcal{L}(\mathbf{x}), \tag{4.6}$$

where $\mathcal{L}(\mathbf{x})$ is the likelihood function defined as

$$\mathcal{L}(\mathbf{x}) = P_N(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} P(y_n|x_n). \tag{4.7}$$

The decoding error occurs when $\mathbf{x}^* \neq \mathbf{x}$. The block code is useful for proving that if $R < \nu_c C$ the probability of decoding error can be made arbitrarily small, i.e. for proving the noisy channel coding theorem:

Theorem 2 (Noisy channel coding theorem)**.** *For every rate $R < \nu_c C$ and $\varepsilon > 0$ there exists N and $\mathcal{S} \subset \mathcal{A}^N$ such that the probability of the decoding error is lower than ε.*

Conversely, if $R > \nu_c C$, there is a lower bound for the probability of the decoding error hence it cannot be made arbitrarily slow.

The proof is given for example in [25, chapter 7].

It is central to this work that the converse part holds for any coding method, not only for the block code, which might not be a biologically relevant decoding strategy.

4.2 Relation to nervous cells

The description of the communication system above illustrates that if there is a source and an information channel satisfying (4.4), it is possible to transmit information from that source reliably with the given channel. In this work, we treat a neuron as an information channel. That said, we are not implying that the communication in the brain follows the scheme described above.

Our goal is to compute the maximal amount of information the neuron could potentially transmit. It is beyond the scope of this work to investigate whether the neuron's maximal potential is employed and if so, how.

4.2.1 Statistical description of neurons

To investigate the information-theoretical properties of neurons as noisy information channels, we need a probabilistic description relating the neuronal input (stimulus) and output (response). We assume that neurons act as a discrete-time memoryless channel without feedback and therefore they are fully described by a conditional probability distribution $P(y|x)$, where x is the stimulus and y is the neuron's response. In this work, we consider the neuron's firing rate - estimated by the number of spikes fired in a given time window - as its response y - i.e. we suppose the neuron employs rate coding described in 3.1.

When a neuron is subjected to a certain stimulus over repeated trials, it is likely to produce a different spike train each time and this can lead to a different number of spikes observed.

Assuming that the neuron's response depends only on the current stimulus, we can assign probabilities to the possible responses (number of fired spikes) given an arbitrary stimulus x by a conditional probability distribution $P(n|x)$, where n is the number of fired spikes. We represent the stimulus x by a number from an interval on the real axis, $x \in [a, b]$, thus reducing the stimulus to the intensity of the stimulus. The dependence of the mean number of spikes fired on the intensity of the stimulus is referred to as the *tuning curve*:

$$\lambda = \sum_{n=0}^{+\infty} nP(n|x) = f(x). \tag{4.8}$$

Example: Poisson neuron

This probabilistic description gives rise to some purely statistical models, describing a neuron's response to a stimulus as a stochastic process. One of the simplest examples is the Poisson neuron. We suppose that a neuron responds to a certain stimulus by firing spikes as a Poisson process with intensity λ, $\lambda \geq 0$. I.e. the neuron fires spikes with time intervals between successive spike drawn from the exponential distribution with probability density function $p(\tau) = \lambda e^{-\lambda\tau}$ and the number of spikes fired in some set time interval ΔT is described by a Poisson random variable N. The probability of observing n spikes in a time interval ΔT is then

$$P(N = n; \lambda, \Delta T) = \frac{(\lambda \Delta T)^n e^{-\lambda \Delta T}}{n!}. \tag{4.9}$$

This probability distribution for several values of intensity λ is shown in the figure 4.2. Note that $E[N] = Var[N] = \lambda$. In the figure 4.3 is shown the conditional probability distribution of a Poisson neuron with a sigmoid tuning curve.

Deterministic models

All the models described in the previous chapters - Hodgkin-Huxley, the integrate and fire models, MAT - are deterministic in the sense, that the response is fully determined by the initial conditions of the model and by the time course of the stimulation (external current or a sequence of synaptic activations).

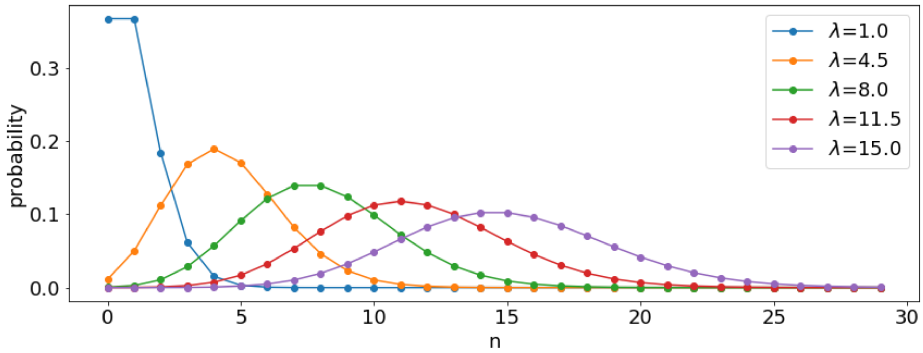

Figure 4.2: The poisson distribution for different values of the intensity λ. The time interval ΔT is considered to be unit.

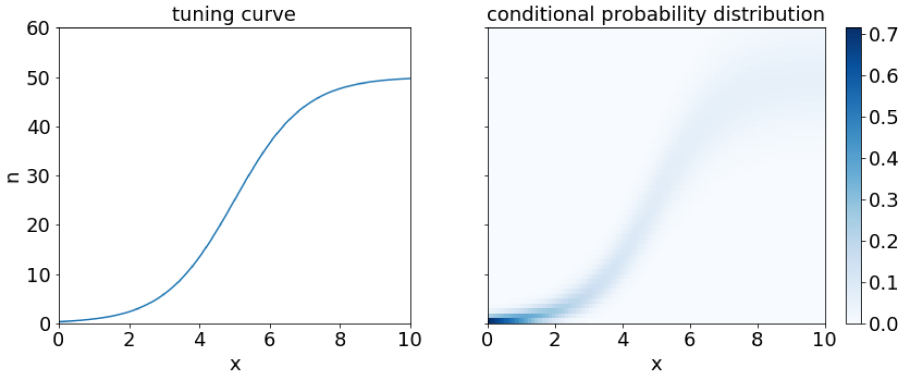

Figure 4.3: The tuning curve $f(x) = \frac{50}{1+e^{-x+5}}$ (left) and the conditional probability distribution of the Poisson neuron (right). The color corresponds to the probability, that for a given stimulus x we will observe n spikes.

However, variability in responses is observed in real neurons. The nervous system is very complex and it cannot be fully described by such simple models. This is typically accounted for by introducing additional stochasticity to the model, e.g. by modeling the input as a stochastic process.

Experiments confirm that when a neuron is stimulated by synapses from many different neurons, modeling the activation times of the synapses as a Poisson process is a good approximation [2]. The intensity of the stimulus can be then interpreted as the intensity of the Poisson process. The conditional probability distribution than has to be obtained by Monte Carlo simulations. Modeling of the input as a Poisson process will be described in more detail in chapter 7.

5. Mutual information and its properties

5.1 Mathematical definition of information capacity

For the purpose of this work we focus on information channels with continuous input alphabet $\mathcal{A} = [a, b]$ and a discrete output alphabet $\mathcal{B} = \{y_1, y_2, ..., y_K\}$ (possibly infinite). However, we will also mention the key differences in channels with a discrete input alphabet or continuous output alphabet.

Suppose we know the conditional probability distribution $P(y|x)$, $x \in \mathcal{A}$, $y \in \mathcal{B}$. We can express the entropy of the output Y given an input symbol x as

$$H(Y|x) = -\sum_{k=1}^{K} P(y_k|x) \log_2 P(y_k|x). \tag{5.1}$$

This is called the *conditional entropy*. If the probability distribution of the input X is known, we can also express the average conditional entropy:

$$H(Y|X) = \int_a^b p(x) H(Y|x) \, \mathrm{d}x, \tag{5.2}$$

where $p(x)$ is a generalized probability density function of the input random variable X (by generalized we mean that it can also contain the Dirac delta functions). Note that in the case of a discrete input alphabet the integral should be replaced by a summation.

$H(Y|X)$ is the part of the response variability, which cannot be accounted for by the changes in stimulus, therefore is also called the noise entropy. The mutual information can then be defined as

$$I(X;Y) = H(Y) - H(Y|X), \tag{5.3}$$

where $H(Y)$ is the Shannon entropy of the marginal output probability distribution q_p:

$$q_p(y) = \int_a^b p(x) P(y|x) \, \mathrm{d}x \quad y \in \mathcal{B}. \tag{5.4}$$

I.e. the mutual information is the uncertainty in response not caused by the noise adherent to the channel.

The formula for mutual information can be rewritten as

$$I(X;Y) = \int_a^b \sum_{k=1}^{K} p(x) P(y_k|x) \log_2 \frac{P(y_k|x)}{\int_a^b p(x') P(y_k|x') \, \mathrm{d}x'} \, \mathrm{d}x. \tag{5.5}$$

This definition of mutual information can also be easily extended to channels with continuous output alphabet:

$$I(X;Y) = \int_{a_{\text{in}}}^{b_{\text{in}}} \int_{a_{\text{out}}}^{b_{\text{out}}} p(x) P(y|x) \log_2 \frac{P(y|x)}{\int_a^b p(x') P(y|x') \, \mathrm{d}x'} \, \mathrm{d}x \, \mathrm{d}y, \tag{5.6}$$

where the intervals $[a_{\text{in}}, b_{\text{in}}]$ and $[a_{\text{out}}, b_{\text{out}}]$ are the input and output alphabets.

It is useful to relate the mutual information to a quantity called the *Kullback-Leibler divergence* or *relative entropy*:

$$D_{KL}(p_1||p_2) = \int_\Omega p_1(x) \log_2 \frac{p_1(x)}{p_2(x)} \, \mathrm{d}x, \tag{5.7}$$

where p_1, p_2 are probability distributions on Ω.

Mutual information is the relative entropy of the joint distribution $p(x, y) = p(x)P(y|x)$ to $p(x)q_p(y)$ (i.e. the case of independence between X and Y):

$$I(X, Y) = D_{KL}(p(x, y)||p(x)q_p(y)). \tag{5.8}$$

It can be shown that the Kullback-Leibler divergence is always non-negative (so called Gibbs inequality), therefore the mutual information is also always non-negative.

Information capacity can be then mathematically defined as the supremum of mutual information over all the possible input probability distributions:

$$C = \sup_{p \in \mathcal{P}([a,b])} I(X; Y), \tag{5.9}$$

where $\mathcal{P}([a, b])$ denotes the space of all generalized probability density functions on the real interval $[a, b]$.

For proof that this quantity is really the information capacity described in chapter 4, see [22, chapters 5, 9] or [25, chapter 7].

5.2 Properties of mutual information and information capacity

5.2.1 Concavity and continuity of information measures

A very useful property of the mutual information is that it is concave as a functional acting upon the probability distribution of the input. This is a key property that guarantees us that if the mutual information is also continuous, there are no local maxima, which is very important when maximizing the mutual information.

Recall that in the case of a discrete output alphabet the mutual information can be expressed as

$$I(X; Y) = H(Y) - H(Y|X), \tag{5.10}$$

where

$$H(Y) = -\sum_{k=1}^{K} q_p(y_k) \log_2 q_p(y_k) \tag{5.11}$$

$$H(Y|X) = -\int_a^b \sum_{k=1}^{K} p(x)P(y_k|x) \log_2 P(y_k|x) \, \mathrm{d}x, \tag{5.12}$$

and $q_p(y)$ is defined by (5.4). It can be seen that the output probability $q_p(y_k)$ and the conditional entropy $H(Y|X)$ are linear in the input distribution p. However, to show the concavity, we need to prove the concavity of $H(Y)$ as a function of q_p.

Lemma 3. *Shannon's entropy $H(Y)$ is a strictly concave function of the probability mass function q of the discrete random variable Y. That is, writing $H(q)$ instead of $H(Y)$, for any two distinct probability mass functions $q^{(1)}$, $q^{(2)}$ and any $\lambda \in (0,1)$:*

$$H(\lambda q^{(1)} + (1-\lambda)q^{(2)}) > \lambda H(q^{(1)}) + (1-\lambda)H(q^{(2)}). \tag{5.13}$$

Proof. Consider any discrete probability distribution q, $q(y_k) = q_k$, a vector $g = (g_1, ..., g_K)$ such that $\sum_{k=1}^{K} g_k = 0$ and $\alpha \in \mathbb{R}$ small enough so that $q_+ \alpha g_k > 0 \; \forall k \in \{1, ..., K\}$. Then

$$H(q + \alpha g) = \sum_{k=1}^{K} (q_k + \alpha g_k) \log_2(q_k + \alpha g_k). \tag{5.14}$$

The derivatives with respect to α are

$$\frac{\mathrm{d}H}{\mathrm{d}\alpha} = -\sum_{k=q}^{K} g_k \log_2(q_k + \alpha g_k) \tag{5.15}$$

$$\frac{\mathrm{d}^2 H}{\mathrm{d}\alpha^2} = -\frac{1}{\ln 2} \sum_{k=1}^{K} \frac{g_k^2}{q_k + \alpha g_k}. \tag{5.16}$$

The second derivative is always negative, therefore H is strictly concave. □

Since q can be viewed as a linear transformation of the input distribution, we can conclude that $H(Y)$ is a concave functional acting upon the input distribution. We will treat $I(X;Y)$ as a map from $\mathcal{P}([a,b])$ to $\mathbb{R}$ and as such we will denote it as $I(p)$, $p \in \mathcal{P}([a,b])$. We can now see from (5.10) that

$$I(X;Y) = I(p) = \text{concave} - \text{linear}, \tag{5.17}$$

therefore the mutual information $I(p)$ is a strictly concave functional.

5.2.2 Attaining capacity, Kuhn-Tucker conditions

The information capacity is defined as the supremum of the mutual information over all possible input probability distributions (5.9). Here we want to show that there exists a unique $p \in \mathcal{P}([a,b])$ such, that the mutual information attains the supremum.

The goal of this part is not the rigorous treatment of the functional $I(p)$, rather to provide a brief explanation for why we can consider that $\sup I(X;Y) = \max I(X;Y)$ and to derive conditions allowing us to identify the optimum.

Central to this part is the notion of weak differentiability and the optimization theorem:

Definition 5.2.1 (Weakly differentiable function)**.** Let Ω be a convex space, $f : \Omega \to \mathbb{R}$, x_0 a fixed element of Ω, $\theta \in [0,1]$. If there exists a map $f'_{x_0} : \Omega \to \mathbb{R}$ such that

$$f'_{x_0}(x) = \lim_{\theta \downarrow 0} \frac{f[(1-\theta)x_0 + \theta x] - f(x_0)}{\theta}, \tag{5.18}$$

than f is said to be weakly differentiable in Ω at x_0 and $f'_{x_0}(x)$ is the weak derivative in Ω at x_0. If f is weakly differentiable in Ω at x_0 for all x_0 in Ω, f is said to be weakly differentiable in Ω.

Theorem 4 (Optimization theorem)**.** *Let f be continuous, weakly differentiable strictly concave map from a compact, convex topological space Ω to $\mathbb{R}$. Define*

$$C = \sup_{x \in \Omega} f(x). \tag{5.19}$$

Then:

1. *$C = \max f(x) = f(x_0)$ for some unique $x_0 \in \Omega$*
2. *A necessary and sufficient condition for $f(x_0) = C$ is $f'_{x_0}(x) \leq 0;\ \forall x \in \Omega$*

Since $\mathcal{P}([a,b])$ is convex and compact (see [26] for exact treatment) and $I(p)$ is a strictly concave map from $\mathcal{P}([a,b])$ to $\mathbb{R}$, we only need to prove that $I(p)$ is also continuous and weakly differentiable.

Continuity of $I(p)$ is generally a property of the channel. It has been shown for several analytically described channels that $I(p)$ is a continuous functional [26], [27], [28]. The proof is always specific for a given conditional probability distribution $P(y|x)$. In most cases in this work we won't know the closed form of $P(y|x)$ -the channel transition probabilities will be obtained by a numerical simulation. Therefore, we cannot prove the continuity. However, it is reasonable to assume that this condition is generally satisfied.

We will define a quantity called the specific information:

$$i(x; q_p) = \sum_{k=1}^{K} P(y_k|x) \log_2 \frac{P(y_k|x)}{q_p(y_k)}, \quad x \in \mathcal{A} \tag{5.20}$$

where q_p is the marginal output probability distribution defined in (5.4). The weak differentiability is then proved by the following lemma.

Lemma 5. *For arbitrary $p_1,\ p_2 \in \mathcal{P}([a,b])$:*

$$I'_{p_1}(p_2) = \int_a^b p_2(x) i(x; q_{p_1})\ \mathrm{d}x - I(p_1). \tag{5.21}$$

Proof.

$$I'_{p_1}(p_2) = \lim_{\theta \downarrow 0} \frac{I[(1-\theta)p_1 + \theta p_2] - I(p_1)}{\theta} \tag{5.22}$$

Denote the output probability mass functions corresponding to the inputs p_1, p_2 as q_1, q_2 respectively:

$$I[(1-\theta)p_1 + \theta p_2] =$$
$$= \int_a^b \sum_{k=1}^{K} [(1-\theta)p_1(x) + \theta p_2(x)] P(y_k|x) \log_2 \frac{P(y_k|x)}{(1-\theta)q_1(y_k) + \theta q_2(y_k)}\ \mathrm{d}x \tag{5.23}$$

$$\log_2 \frac{P(y_k|x)}{(1-\theta)q_1(y_k) + \theta q_2(y_k)} = \tag{5.24}$$
$$= \log_2 \frac{P(y_k|x)}{q_1(y_k)} - \log_2 \left(1 + \theta \left(\frac{q_2(y_k)}{q_1(y_k)} - 1\right)\right) \tag{5.25}$$
$$\approx \log_2 \frac{P(y_k|x)}{q_1(y_k)} - \theta \left(\frac{q_2(y_k)}{q_1(y_k)} - 1\right) \tag{5.26}$$

By putting (5.26) into (5.23) and neglecting the higher order terms we obtain:

$$I[(1-\theta)p_1+\theta p_2] \approx \int_a^b \sum_{j=1}^{J}\Bigg[(1-\theta)p_1(x)P(y_k|x)\log_2\frac{P(y_k|x)}{q_1(y_k)}+ \\ +\theta p_2(x)\log_2\frac{P(y_k|x)}{q_1(y_k)}- \\ -p_1(x)\theta\left(\frac{q_2(y_k)}{q_1(y_k)}-1\right)\Bigg]\,\mathrm{d}x+ = \\ =(1-\theta)I(p_1)+\theta\int_a^b p_2(x)i(x;q_{p_1})\,\mathrm{d}x \quad (5.27)$$

Finally, the proof is completed by putting (5.27 into 5.22). □

We can now conclude that the optimal input probability distribution exists and is unique.

Using (5.21) and (5.20), we can rewrite the 2. statement of the Theorem 4 as

$$\int_a^b p(x)i(x;q_{p_0})\,\mathrm{d}x \le I(p_0). \quad (5.28)$$

This leads to the following formulation of so called Kuhn-Tucker conditions:

Theorem 6 (Kuhn-Tucker conditions)**.** *Let p_0 be an arbitrary generalized probability density function on $[a,b]$. Let E_0 denote the support of p_0. Then p_0 is optimal if and only if*

$$i(x;q_{p_0}) \le I(p_0) \quad \forall x \in [a,b], \quad (5.29)$$
$$i(x;q_{p_0}) = I(p_0) \quad \forall x \in E_0. \quad (5.30)$$

We will follow the proof in [26].

Proof. If both conditions hold, the necessary and sufficient condition for the maximum from the Theorem 4 is satisfied and therefore p_0 is optimal. This proves $\Leftarrow$.

To prove $\Rightarrow$, first suppose that p_0 is optimal, but doesn't satisfy the first condition. Then there exists $x_1 \in [a,b]$ such that $i(x_1;q_{p_0}) > I(p_0)$ and for $p(x) = \delta(x-x_1)$, where δ denotes the Dirac delta function, holds

$$\int_a^b p(x)i(x;q_{p_0})\,\mathrm{d}x = i(x_1;q_{p_0}) > I(p_0). \quad (5.31)$$

This contradicts (5.28).

Now suppose that p_0 is optimal, but doesn't satisfy the second condition. Because the first statement is valid, a subset $E' \subset E_0$ with positive measure (i.e. $\int_{E'} p(x)\,\mathrm{d}x = \varepsilon > 0$) and such, that $i(x;q_{p_0}) < I(p_0)\ \forall x \in E'$ must exist. Then clearly $I(p_0) < I(p_0)$ which is a contradiction. □

Constrained channels

In some cases we want to search for the optimal solution in a subset of $\mathcal{P}([a,b])$. Typically, we can assign expenses to individual symbols in the input alphabet. We shall denote the function assigning the expenses as e, $e: \mathcal{A} \to \mathbb{R}$. We can then require the average expense not to exceed some given limit E, i.e. $\int_a^b p(x)e(x)\ \mathrm{d}x \leq E$.

For these cases we need the Lagrangian theorem [29].

Theorem 7 (Lagrangian Theorem)**.** *Let Ω be a convex metric space, f and g concave functionals on Ω to $\mathbb{R}$. Assume that there exists an $x_1 \in \Omega$ such that $g(x_1) < 0$ and let*

$$C' = \sup_{\substack{x\in\Omega \\ g(x)\leq 0}} . \tag{5.32}$$

If C' is finite, then there exists a constant $\mu > 0$ such that

$$C' = \sup_{x\in\Omega}[f(x) - \mu g(x)]. \tag{5.33}$$

Furthermore, if the supremum in the first equation is achieved by x_0 in Ω and $g(x_0) \leq 0$, it is achieved by x_0 in the second equation, and $\mu g(x_0) = 0$.

In the case of the average expense limitation, we will set

$$g(p) = \int_a^b p(x)e(x)\ \mathrm{d}x - E. \tag{5.34}$$

The functional $g(p)$ is linear in $p(x)$, therefore it is concave (not strictly) and the functional

$$I_E(p) = I(p) - g(p) \tag{5.35}$$

is strictly concave. This implies that for any $E > 0$ there exists a corresponding unique multiplier μ and there exists a unique solution of (5.33) (from the Theorem 4).

Since

$$g'_{p_1}(p_2) = g(p_2) - g(p_1), \tag{5.36}$$

the second property of the Theorem 4 can be then written as

$$\int_a^b p(x)[i(x;q_{p_0}) - \mu e(x)]\ \mathrm{d}x \leq I(p_0) - \mu E \tag{5.37}$$

and the Kuhn-Tucker conditions can be modified accordingly.

Theorem 8 (Kuhn-Tucker conditions for a constrained channel)**.** *Let p_0 be an arbitrary generalized probability density function on $[a,b]$. Let E_0 denote the support of p_0. Then p_0 is optimal if and only if*

$$i(x;q_{p_0}) - \mu e(x) \leq I(p_0) - \mu E \quad \forall x \in [a,b], \tag{5.38}$$

$$i(x;q_{p_0}) - \mu e(x) = I(p_0) - \mu E \quad \forall x \in E_0. \tag{5.39}$$

Finite alphabet

So far we've only discussed the case when the input alphabet is a real interval. However, on the computer we have to consider only a finite number of input symbols. For this reason we will now briefly discuss the case of a finite input alphabet of size J, $\mathcal{A} = \{x_1, ..., x_J\}$.

Let's consider a concave multivariate function $f(p)$, $p = (p_1, ..., p_J)$, $f \in C^1(\mathbb{R}^J)$. It is a fundamental result of the multivariate calculus, that such function will attain its maximum on a compact set. The set of our interest will be that of non-negative vectors of length J summing to 1, i.e. set of vectors representing discrete probability distributions on the given finite alphabet:

$$\mathcal{P}^J = \{p \in \mathbb{R}^J \mid \sum_j \alpha_j = 1,\ \alpha_j \geq 0\ \forall j\}. \tag{5.40}$$

$\mathcal{P}^J$ is a closed subspace of a finite dimensional space, therefore it is compact.

We are interested in maximizing the mutual information defined as

$$I(X;Y) = \sum_{j=1}^{J}\sum_{k=1}^{K} p_j P(y_k|x_j) \log_2 \frac{P(y_k|x_j)}{\sum_{l=1}^{J} p(x_l)P(y_k|x_l)} \tag{5.41}$$

as a function of the probability assignments $p_j = p(x_j)$ to the input alphabet. The argumentation that $I(X;Y)$ is concave from 5.2.2 still applies, therefore we can conclude that there exists a unique vector of probability assignments maximizing the mutual information.

Next we want to write the Kuhn-Tucker conditions for the case of a finite input alphabet. Using the Lagrange multiplier method we obtain the following Lagrangian function:

$$\mathcal{L}(p_1, ..., p_J, \lambda) = I(X;Y) - \lambda(1 - \sum_{j=1}^{J} p_j). \tag{5.42}$$

The optimality conditions on the maximizer p_0 are then:

$$\begin{aligned} \frac{\partial I(X;Y)}{\partial p_k}|_{p_0} &= \lambda \quad \forall k:\ p_k > 0 \\ \frac{\partial I(X;Y)}{\partial p_k}|_{p_0} &\leq \lambda \quad \forall k:\ p_k = 0. \end{aligned} \tag{5.43}$$

The second condition considers the additional constraint $p_j \geq 0\ \forall j$, i.e. sets the condition for the situation when the vector is on the border of the set. This is discussed in more detail in [22, chapter 4].

The partial derivatives of (5.41) take on the following form:

$$\frac{\partial I(X;Y)}{\partial p_l} = \sum_{k=1}^{K} P(y_k|x_l) \log_2 \frac{P(y_k|x_l)}{q_p(y_k)} - \log_2 e = i(x_l; q_p) - \log_2 e, \tag{5.44}$$

where $i(x_l; q_p)$ is the specific information defined by (5.20).

Using the conditions (5.43) and denoting $C = \lambda + \log_2 e$ we obtain the discrete form of the Kuhn-Tucker conditions for the optimal probability distribution p_0:

$$i(x_j; q_{p_0}) = C \quad \forall j : \ p_0(x_j) > 0 \tag{5.45}$$

$$i(x_j; q_{p_0}) \leq C \quad \forall j : \ p_0(x_j) = 0. \tag{5.46}$$

From (5.45) it is obvious that C really has the meaning of capacity.

The case of an expense constrained channel is nearly identical. We consider a vector $e = (e_1, ..., e_J)$, where e_j is the expense associated with the input symbol x_j. The space of vectors representing the probability distributions satisfying the constraint

$$\mathcal{P}_E^J = \{p \in \mathcal{P}^J | \sum_{j=1}^{J} p_j e_j \leq E\} \tag{5.47}$$

is still compact and therefore there exists a unique $p_0 \in \mathcal{P}_E^J$ maximizing the mutual information. Using a Lagrange multiplier μ, we can modify the Kuhn-Tucker conditions:

$$i(x_j; q_{p_0}) - \mu e_j = V \quad \forall j : \ p_0(x_j) > 0 \tag{5.48}$$

$$i(x_j; q_{p_0}) - \mu e_j \leq V \quad \forall j : \ p_0(x_j) = 0. \tag{5.49}$$

where $V = C - \mu E$.

5.3 Examples of analytically solvable information channels

In special cases it is possible to find a closed form solution for the information capacity and the optimal input distribution. In this section are some of these cases described.

Gallager's phase channel

[22, Exercise 7.4] The input and output alphabet of this channel are real numbers from the interval $[0, 2\pi)$. The channel is subjected to an additive noise Z independent of the input and is described by a probability density function $p_Z(z)$, which is non-zero only for $z \in [0, 2\pi)$. The output Y of the channel is then given by

$$Y = (X + Z) \bmod 2\pi, \tag{5.50}$$

where X is the input and Z is a continuous random variable specified by the probability density function p_Z.

Since all inputs to this channel are equivalent, it is easy to see that the Kuhn-Tucker conditions (5.29), (5.30) are satisfied for the uniform probability density function of the inputs $p_X(x) = \frac{1}{2\pi}$ and all the outputs are then equally probable ($p_Y(y) = \frac{1}{2\pi}$). Therefore, the specific information for any input x is given by

$$i(x; \frac{1}{2\pi}) = \int_0^{2\pi} p_Z(z) \log_2 \frac{p_Z(z)}{\frac{1}{2\pi}} = I(X;Y) = C. \tag{5.51}$$

Possible example is $p_Z(z) = \frac{1}{\alpha}$ for $z \in [0, \alpha)$ and 0 otherwise. The integral is then trivial and $C = \log_2 \frac{2\pi}{\alpha}$.

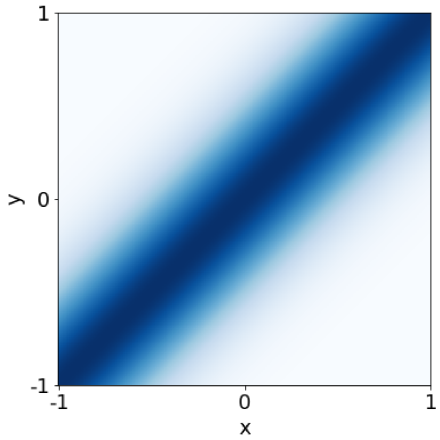

Figure 5.1: Transition probabilities $p(y|x)$ for the Gaussian channel. The color represents the probability density function. The darker the color, the higher probability density of y for a given x.

Gaussian channel

Similarly as in the case of Gallager's phase channel, the channel is subjected to an additive noise, independent of the input. However, the input and output alphabets are all real numbers. The channel's response Y to input X is

$$Y = X + Z, \tag{5.52}$$

where Z is a normally distributed random variable with a zero mean and a variance σ^2. The transition probabilities are visualized in the figure 5.1.

For this case a power constraint

$$\mathrm{Var}(X) \leq \mathcal{E} \tag{5.53}$$

is usually considered. Given such a constraint the capacity is reached when the probability distribution of X is the normal distribution with variance $\mathcal{E}$. The mutual information (and information capacity) is then:

$$C = \frac{1}{2}\log_2\left(1 + \frac{\mathcal{E}}{\sigma^2}\right). \tag{5.54}$$

For details see [22, Chapter 7].

Discrete channel with M inputs and M outputs

For a subclass of channels with discrete input and output alphabets it is possible to find a closed form solution for the optimal input probability distribution and capacity. These channels have to satisfy the following:

- The transition probability matrix P, $P_{ij} = P(y_i|x_j)$ is invertible. It is necessary (but not satisfactory), that the size of input and output alphabet is the same, i.e. $|\mathcal{A}| = |\mathcal{B}| = M$

- The optimal input probability distribution p is non-zero for all symbols in the input alphabet

In such case the optimal probabilities are given by

$$p(x_k) = \exp(\lambda - 1) \sum_{j=1}^{M} q_{jk} \exp\left[- \sum_{i=1}^{M} q_{ji} H(Y|x_i) \right] \tag{5.55}$$

$$1 - \lambda = \log \sum_{j=1}^{M} \exp\left[- \sum_{i=1}^{M} q_{ji} H(Y|x_i) \right], \tag{5.56}$$

where q_{ji} are elements of $Q = P^{-1}$. The condition on non-zero probability for each letter ensures, that the probability assignments $p(x_k)$ computed by this formula are positive. The capacity is then

$$C = \log_2 \sum_{j=1}^{M} \exp_2\left[- \sum_{i=1}^{M} q_{ji} H(Y|x_i) \right]. \tag{5.57}$$

The proof of this formula is provided in [24, Chapter 3]

6. Numerical methods for mutual information optimization

6.1 Blahut-Arimoto algorithm

Above we showed some examples of analytically solvable information channels. However, in most cases the closed form of the capacity and the optimal input probability distribution is not known and has to be evaluated numerically or approximately.

We have established above that finding the optimal input distribution is a convex problem. If we are searching for an optimal input distribution of a channel with a finite input and output alphabet, any convex optimization method can be used (e.g. interior point methods [30]). In this section, however, we will describe an algorithm that was developed specifically for maximizing mutual information and is often more practical for this purpose than a universal optimization algorithm.

6.1.1 Capacity of unconstrained discrete channels

The Blahut-Arimoto algorithm is based on the following theorem [31]

Theorem 9. *Suppose the channel transition matrix Q, $Q_{k|j} = P(y_j|x_k)$ is $n \times m$. For any $m \times n$ transition matrix P, let*

$$J(p, Q, P) = \sum_j \sum_k p_j Q_{k|j} \log \frac{P_{j|k}}{p_j}, \tag{6.1}$$

where $p_j = p(x_j)$ are input probability assignments. Then the following is true:

a) $C = \max_p \max_P J(p, Q, P)$

b) For fixed P, $J(p, Q, P)$ is maximized by

$$P_{j|k} = \frac{p_j Q_{k|j}}{\sum_j p_j Q_{k|j}}. \tag{6.2}$$

c) For fixed P, $J(p, Q, P)$ is maximized by

$$p_j = \frac{\exp\left(\sum_k Q_{k|j} \log P_{j|k}\right)}{\sum_l \exp\left(\sum_k Q_{k|l} \log P_{l|k}\right)} \tag{6.3}$$

By employing the theorem it is possible to construct a sequence of input probability distributions in the following manner:

1. Choose any vector p_0, representing the initial guess at the optimal input probability distribution. Set $p = p_0$.

2. For the fixed p, maximize J by setting $P_{j|k} = \frac{p_j Q_{k|j}}{\sum_j p_j Q_{k|j}}$.

3. Fix newly obtained transition matrix $P_{j|k}$ and maximize J by setting $p_j = \frac{\exp\left(\sum_k Q_{k|j} \log P_{j|k}\right)}{\sum_l \exp\left(\sum_k Q_{k|l} \log P_{l|k}\right)}$. Continue with 2.

The steps 2. and 3. can be done together and thus generate a sequence converging to the optimum.

Theorem 10. *For any $p \in \mathcal{P}^J$, let*

$$c_j(p) = \exp\left(\sum_k Q_{k|j} \log \frac{Q_{k|j}}{\sum_l p_l Q_{k|l}}\right) \tag{6.4}$$

$$P^*_{j|k} = \frac{p_j Q_{k|j}}{\sum_j p_j Q_{k|j}} \tag{6.5}$$

$$I(p, Q) = \sum_j \sum_k p_j Q_{k|j} \log \frac{P^*_{j|k}}{p_j} \tag{6.6}$$

Then, if p^0 is any element of $\mathcal{P}^J$ with all components strictly positive, the sequence of probability vectors defined by

$$p_j^{r+1} = p_j^r \frac{c_j^r}{\sum_l p_l^r c_l(p^r)} \tag{6.7}$$

is such that $I(p^r, Q) \to C$ as $r \to +\infty$.

It is clear that $\{I(p^r, Q)\}_r$ will be a non-decreasing sequence bounded from above by the capacity C and therefore has to converge to some I^∞. Moreover, by the Bolzano-Weierstrass theorem the sequence $\{p_r\}_r$ has to have a limit point p^*. What remains to prove is that for this limit point the Kuhn-Tucker conditions (eqs. (5.45), (5.46)) are satisfied and therefore $I^\infty = C$. For the complete proof see the original article [31].

6.1.2 Capacity of constrained discrete channels

The extension of the Blahut-Arimoto algorithm to channels with the constraint

$$\sum_j p_j e_j \leq E \tag{6.8}$$

is also provided in [31]. The goal is to maximize the quantity

$$\sum_{j=1}^{J} p_j \left[i(x; q_p) - \mu e_j\right], \tag{6.9}$$

where μ is such that for the optimal probability distribution p^* the condition (6.8) holds. The constrained capacity $C(E)$ is then

$$C(E) = \sum_{j=1}^{J} p_j^* i(x_j; q_{p^*}) \tag{6.10}$$

For an arbitrary value of the multiplier $\mu \in [0, +\infty]$ one can use the following generalization of the Theorem 10:

Theorem 11. *Let $\mu \in [0, +\infty]$ be given and for any $p \in \mathcal{P}^J$ and a channel transition matrix Q, $Q_{kj} = P(y_k|x_j)$ let*

$$c_j(p) = \exp\left(\sum_k Q_{kj} \log \frac{Q_{kj}}{\sum_l p_l Q_{kl}} - \mu e_j\right). \tag{6.11}$$

Then, if p^0 is any element of $\mathcal{P}^J$ with all components strictly positive, the sequence of probability vectors defined by

$$p_j^{r+1} = p_j^r \frac{c_j^r}{\sum_l p_l^r c_l(p^r)} \tag{6.12}$$

is such that

$$\sum_{j=1}^{J} p_j e_j \to E_\mu \qquad \text{as} \quad r \to +\infty \tag{6.13}$$

$$I(p^r, Q) \to C(E_\mu) \qquad \text{as} \quad r \to +\infty. \tag{6.14}$$

By constructing the sequence $\{p^n\}$ for a fixed μ we obtain $C(E_\mu)$ for some expense E_μ.

If the value of the multiplier μ is not known beforehand (as is usually the case), the following lemma can be utilized to find the constrained capacity for a given expense E:

Lemma 12.

$$C(E) = \min_{s \in [0, +\infty]} C(E_s) \tag{6.15}$$

The proof is provided in [31]

6.2 Bounds to the information capacity

While maximizing the mutual information with an arbitrary optimization algorithm, one can easily control the precision of the solution. This is formulated in the following lemma:

Lemma 13. *For an arbitrary probability distribution $p \in \mathcal{P}([a, b])$ and a channel with transition probabilities $P(y|x)$ and a capacity C, the following holds:*

$$\int_a^b p(x) i(x; q_p)\, \mathrm{d}x \leq C \leq \max_{x \in [a,b]} i(x; q_p). \tag{6.16}$$

Proof. The first inequality is clear - the maximal mutual information cannot be lower than the mutual information corresponding to any arbitrary probability assignment.

To prove the second inequality, consider an arbitrary input probability distribution $p_1 \in \mathcal{P}([a, b])$ and a corresponding output probability distribution

$q_{p_1}(y) = \int_a^b p_1(x)P(y|x)\ \mathrm{d}x,\ y \in \mathcal{B}$. We employ the inequality $\ln x \geq 1 - \frac{1}{x}$ with equality if and only if $x = 1$:

$$\int_a^b \sum_{k=1}^{K} p(x)P(y_k|x) \log_2 \frac{q_{p_1}(y_k)}{q_p(y_k)}\ \mathrm{d}x \geq$$
$$\geq \int_a^b \sum_{k=1}^{K} p(x)P(y_k|x) \log_2 e \left(1 - \frac{q_p(y_k)}{q_{p_1}(y_k)}\right)\ \mathrm{d}x = 0 \quad (6.17)$$

Therefore,

$$\int_a^b \sum_{k=1}^{K} p(x)P(y_k|x) \log_2 \frac{P(y_k|x)}{q_p(y_k)}\ \mathrm{d}x \geq$$
$$\geq \int_a^b \sum_{k=1}^{K} p(x)P(y_k|x) \log_2 \frac{P(y_k|x)}{q_{p_1}(y_k)}\ \mathrm{d}x \quad (6.18)$$

with equality if and only if $q_p = q_{p_1}\ \ \forall y \in \mathcal{B}$.

Rewriting the inequality (6.18) we obtain

$$I(p) \leq \int_{\mathcal{A}} p(x)i(x; q_{p_1})\ \mathrm{d}x \leq \max_{x \in [a,b]} i(x; q_{p_1}) \quad (6.19)$$

□

The proof holds just as well for discrete input alphabets, one only needs to replace the integration with a summation.

Therefore, if we maximize the mutual information iteratively, we can compute the upper bound in each step (the lower bound is the current value of mutual information) and decide if the precision in sufficient.

6.3 Optimization for continuous input channels

The Blahut-Arimoto algorithm serves for computing the capacity of channels with discrete input and output alphabets. However, in neurons we usually consider the input to be a stimulus intensity represented by a number from an interval on the real axis. We will discuss two strategies of dealing with this problem - the extension of the Blahut-Arimoto algorithm to continuous input alphabets and methods based on the assumption, that the support of the capacity-achieving distribution has only a finite number of points (we will call those methods *alphabet optimization methods*).

6.3.1 Continuous Blahut-Arimoto

Blahut proposes that substituting the corresponding summations in the algorithm by integrals makes the algorithm suitable for channels with a continuous input alphabet [31]. By employing this approach the sequence defined by the eqs. (6.6),

(6.7) becomes

$$c(x;p) = \exp\left(\sum_k Q_{k|j} \log \frac{Q_{k|j}}{\int_a^b p(x)P(y_k|x)\ \mathrm{d}x}\right) \tag{6.20}$$

$$p^{r+1}(x) = p^r(x)\frac{c(x;p^r)}{\int_a^b p^r(x)c(x;p^r)} \tag{6.21}$$

In the case where the conditional probability distribution $P(y|x)$ is obtained by a Monte Carlo simulations, the distribution is usually available only for a finite number of stimulus intensities x. The integral then can be evaluated by standard methods of numerical integration, e.g. the Newton-Cotes formulas. However, these methods require the integrated function to be smooth, which (as we will argue later) is not usually the case for (generalized) input probability density functions corresponding to near-capacity mutual information values. Therefore, it is best to use the simplest method - the rectangular rule. But this essentially means treating the channel as a discrete one and using the classical Blahut-Arimoto algorithms for discrete channels.

6.3.2 Alphabet optimization

It has been shown for several different information channel models that the support of the optimal input probability distribution has only a finite number of points [26], [28]. Moreover, if the output alphabet is finite, i.e. $|\mathcal{B}| = K$, the capacity can be achieved by an input distribution with only K or fewer points of support [32].

We propose a simple iterative optimization procedure for constrained channels, which is a generalization of [33, Algorithm 2] to constrained channels. Our goal is to compute $C(E_\mu)$, where E_μ is the expense parameterized by a fixed multiplier μ. Define

$$v(x;q_p) = \sum_{k=1}^{N} P(y_k|x)\frac{P(y_k|x)}{q_p(y_k)} - \mu e(x), \tag{6.22}$$

where

$$q_p = \int_a^b p(x)P(y_k|x)\ \mathrm{d}x. \tag{6.23}$$

Then

$$C(E_\mu) = \mu E_\mu + \sup_{p\in\mathcal{P}([a,b])} \int_a^b p(x)v(x;q_p)\ \mathrm{d}x. \tag{6.24}$$

Therefore, our goal is to find such input alphabet $\mathcal{A}*$, that

$$\max_{p\in\mathcal{P}([a,b])} \int_a^b p(x)v(x;q_p)\ \mathrm{d}x = \max_{p\in\mathcal{P}(\mathcal{A}^*)} \sum_{x\in\mathcal{A}^*} p(x)v(x;q_p) = V \tag{6.25}$$

By the following algorithm we will construct a sequence of input alphabets $\{\mathcal{A}^n\}$, that the sequence $\{V^n\}$ defined by

$$V^n = \sum_{x\in\mathcal{A}^n} p^n(x)v(x;q_{p^n}), \tag{6.26}$$

where p^n is the optimal input distribution on the input alphabet $\mathcal{A}^n$, converges to V.

Algorithm

1. Chose an arbitrary initial input alphabet $\mathcal{A}^0 = x_1^0, ..., x_J^0$ (e.g. the two extreme point of the input alphabet). Set $n = 0$

2. Set $n = n + 1$. Use the constrained Blahut-Arimoto algorithm to find the optimal distribution for the alphabet $\mathcal{A}^n$. Obtain:

   - V^n
   - The optimal input distribution p^n on the input alphabet $\mathcal{A}^n$
   - The output probability vector $q_{p^n}(y) = (q_{p^n}(y_1), ..., q_{p^n}(y_K))$, where $q_{p^n}(y_k) = \sum_{x \in \mathcal{A}^n} p^n(x) P(y_k|x)$

3. Define
$$\epsilon_n = \max_{x \in [a,b]} v(x; q_{p^n}) - V^n. \tag{6.27}$$
   Terminate if the desired precision ϵ has been reached (i.e. $\epsilon_n < \epsilon$). Otherwise continue.

4. Create a new input alphabet $\mathcal{A}^{n+1}$ with all zero-probability symbols removed and symbols locally maximizing the function $v(x; q_{p^n})$ added.

5. Continue with 2.

Theorem 14. *The sequence $\{V^n\}$ is non-decreasing and $V^n = V^{n+1}$ if and only if $V^n = \max_{x \in [a,b]} v(x; q_{p^n})$.*

Proof. In the $(n+1)$th step of the algorithm we can define the following probability distribution on $\mathcal{A}^{n+1}$:
$$\tilde{p}^{n+1}(x) = \begin{cases} p^n, & x \in \mathcal{A}^{n+1} \cap \mathcal{A}^n \\ 0, & x \in \mathcal{A}^{n+1} \setminus \mathcal{A}^n \end{cases} \tag{6.28}$$

Then
$$\sum_{x \in \mathcal{A}^n} p^n(x) v(x; q_{p^n}) = \sum_{x \in \mathcal{A}^{n+1}} \tilde{p}^{n+1}(x) v(x; q_{\tilde{p}^{n+1}}) \leq \sum_{x \in \mathcal{A}^{n+1}} p^{n+1}(x) v(x; q_{p^{n+1}}) \tag{6.29}$$
This proves that the sequence is non-decreasing.

$$\underline{V^n = V^{n+1} \Rightarrow V^n = \max_{x \in [a,b]} v(x; q_{p^n}) :}$$

From lemma 13, we know that
$$V^{n+1} \geq \sum_{x \in \mathcal{A}^{n+1}} \tilde{p}^{n+1}(x) v(x; q_{\tilde{p}^{n+1}}) = \sum_{x \in \mathcal{A}^n} p^n(x) v(x; q_{p^n}) = V^n. \tag{6.30}$$

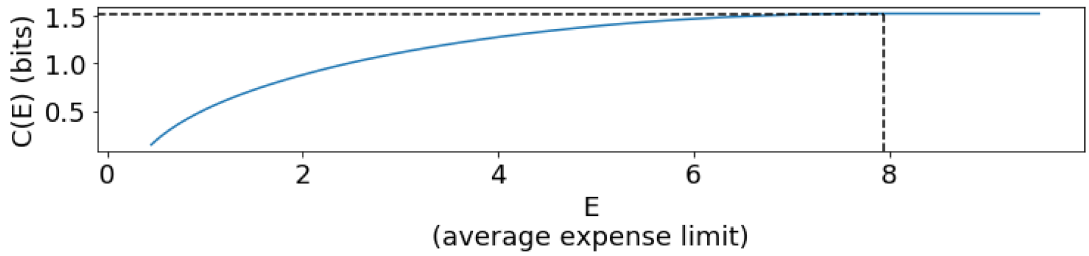

Figure 6.1: Information capacity of the Poisson neuron as a function of the average expense limit. The dashed lines denote the point from which on the constraint play no longer any role.

If $V^{n+1} = V^n$, then $p^{n+1} = \tilde{p}^{n+1}$. Therefore

$$V^{n+1} = \max_{x \in \mathcal{A}^{n+1}} v(x; q_{\tilde{p}^{n+1}}) \tag{6.31}$$

$$\geq \max_{x \in [a,b]} v(x; q_{p^n}) \geq V^n = V^{n+1}. \tag{6.32}$$

Subsequently $V^n = \max_{x \in [a,b]} v(x; q_{p^n})$.

$$\underline{V^n = \max_{x \in [a,b]} v(x; q_{p^n}) \Rightarrow V^n = V^{n+1}:}$$

Now assume $V^n = \max_{x \in [a,b]} v(x; q_{p^n})$. Then

$$V^{n+1} = \max_{x \in \mathcal{A}^{n+1}} v(x; q_{p^{n+1}}) \tag{6.33}$$

$$\leq \max_{x \in \mathcal{A}^{n+1}} v(x; q_{p^n}) \tag{6.34}$$

$$\leq \max_{x \in \mathcal{A}^n} v(x; q_{p^n}) = C^n \tag{6.35}$$

Therefore, $V^n \geq V^{n+1}$. But the sequence $\{V^n\}$ is non-decreasing, implying that $V^{n+1} = V^n$

□

6.4 Information capacity of the Poisson neuron

We applied the alphabet optimization method to evaluate the information capacity of the Poisson neuron from the figure 4.3. We set a constraint to the average number of spikes fired, i.e. the expense function was

$$e(x) = \sum_{k=0}^{+\infty} kP(k|x) = f(x), \tag{6.36}$$

where $f(x)$ is the tuning curve as defined in (4.8).

Thus, we are searching for the constrained information capacity $C(E)$ given that the average count of output spikes does not exceed some limit E. A function describing the dependence of $C(E)$ on E is called a capacity-cost function. In the figure 6.1 the capacity-cost function for the Poisson neuron is depicted. At one point the capacity becomes constant and corresponds to the unconstrained problem. In the figure 6.2 is showed in detail how the alphabet optimization method converges.

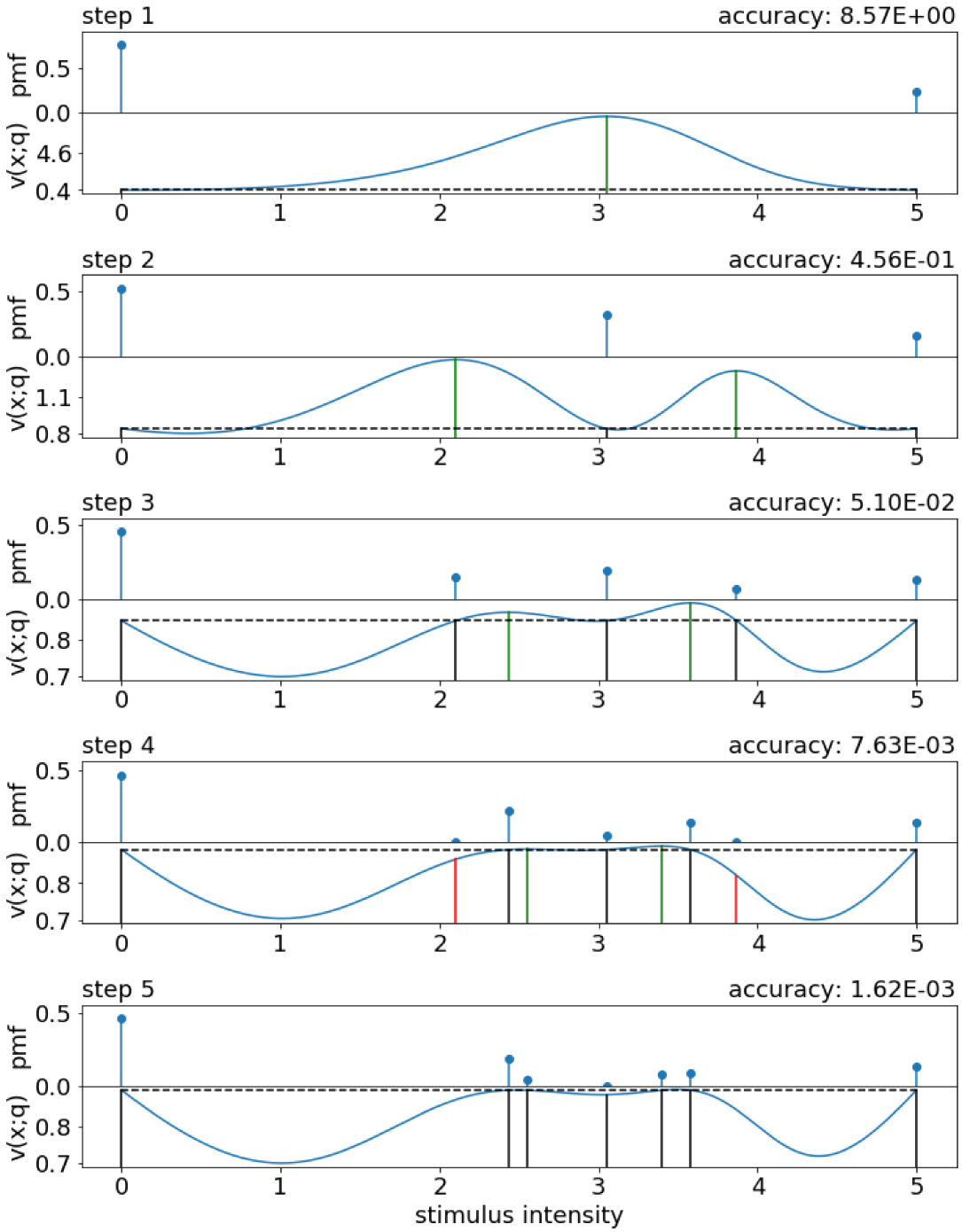

Figure 6.2: Convergence of the alphabet optimization method for the Poisson neuron with $\mu = 0.1$ corresponding to about 3 spikes on the output on average. The upper plot shows the probability distribution p, the lower plot shows the value $v(x; q_p)$. Black vertical lines correspond to the current input alphabet $\mathcal{A}^n$, the green are to be added to the next steps and the red are to be removed. The dashed horizontal line is the current value of V^n, showing that the Kuhn-Tucker conditions (5.48, 5.49) are satisfied for the finite alphabet $\mathcal{A}^n$. The accuracy is the difference between the maximum of $v(x; q_p)$ and the dashed line.

7. Information capacity of the MAT neuron

In 4.2.1 we described how it is possible to describe a neuron as an information channel. In this chapter we will go through the details and use the results to compute the information capacity of the MAT model and how it depends on the model's parameters and external constraints.

To evaluate the capacity of a deterministic neuron model we will follow these steps (we will focus on the MAT model, but the points below are sufficiently general):

- Describing the stimulus (neural input) ensemble and modeling the input
- Modeling of the neuron's response to an input from the ensemble
- Sampling the distribution $P(y|x)$ for multiple stimulus intensities x.
- Using the conditional probability distribution $P(y|x)$ to evaluate the information capacity.

7.1 Input generation and synaptic conducances integration

7.1.1 Input generation

We already mentioned in 4.2.1 that modeling the activation of synapses as a Poisson process is a good approximation of the reality. We also described some basic properties of the Poisson process with intensity λ, namely that the probability density function of T - time intervals between subsequent events - is exponential:

$$p(T = \tau; \lambda) = \lambda e^{-\lambda\tau} \tag{7.1}$$

and the number of events N that occur in a time interval Δt obeys the Poisson distribution:

$$P(N = n; \lambda, \Delta t) = \frac{(\lambda\Delta t)^n e^{-\lambda\Delta t}}{n!}. \tag{7.2}$$

The intensity of the process λ can be roughly interpreted as the intensity of the stimulus, but the actual definition of stimulus intensity will depend on the input generation method used. Note that a neuron typically has both excitatory and inhibitory synapses. In our simulations we will also consider both types of the synapses.

7.1.2 Synaptic conductances integration

The equations (1.20) and (1.21) describe how the total current can be computed. From computational point of view it is very inconvenient to remember all times of

past synaptic activations and evaluating these formulas in each time step would be very time-consuming.

However, it is possible to reformulate them in such a way that greatly simplifies the computation. When no synapses occur in time interval $(t, t + \Delta t)$, following holds:

$$g_{\mathrm{AMPA}}(t + \Delta t) = g_{\mathrm{AMPA}}(t) e^{-\Delta t / \tau_{\mathrm{AMPA}}}. \tag{7.3}$$

Occurrence of a synapse at time t can be modeled as

$$g_{\mathrm{AMPA}}(t) \to g_{\mathrm{AMPA}}(t) + P_{\max} \bar{g}_{\mathrm{AMPA}}. \tag{7.4}$$

When we move forward in discrete time steps Δt, equations (7.3) and (7.4) can be combined together, and we obtain the following relation:

$$(g_{\mathrm{AMPA}})_{n+1} = (g_{\mathrm{AMPA}})_n e^{-\Delta t / \tau_{\mathrm{AMPA}}} + k P_{\max} \bar{g}_{\mathrm{AMPA}}, \tag{7.5}$$

where $k \in \{0, 1, 2, ...\}$ is the number of synapses that occurred in time interval $(t, t + \Delta t)$, $(g_{\mathrm{AMPA}})_n$ is the approximation of g_{AMPA} in time $t_n = n\Delta t$.

Finally, the approximation of synaptic current $(i_{\mathrm{AMPA}})_n$ is then obtained as

$$(i_{\mathrm{AMPA}})_n = V_n (g_{\mathrm{AMPA}})_n. \tag{7.6}$$

7.1.3 Current injection

Sometimes it is convenient to simulate direct current injection instead of synaptic conductances. It is much easier to achieve this kind of input in an experimental setting which allows for an easier comparison between an experiment and a simulation. This approach was used in the work introducing the MAT model, therefore we will also employ it in our simulations [5].

The input current is

$$I(t) = A \sum_k I_{\mathrm{exc}} g_{\tau_{\mathrm{exc}}}(t - t_k) + A \sum_j I_{\mathrm{inh}} g_{\tau_{\mathrm{inh}}}(t - t_j), \tag{7.7}$$

where

$$g_\tau(t) = \begin{cases} \frac{t}{\tau} e^{-t/\tau}) & t \geq 0 \\ 0 & t < 0. \end{cases} \tag{7.8}$$

A, I_{exc}, I_{inh}, τ_{exc} and τ_{inh} are constants, $\{t_k\}$, $\{t_j\}$ are times the interval between which are drawn from the exponential distribution (similarly as in the case of synaptic input). In this manner we discard the dependence of the current on the potential difference.

7.1.4 Used parameters

Current injection

In the case of the current injection, we used the same parameters as in the original paper [5]:

$$\begin{aligned} I_{\mathrm{exc}} &= 0.1\,\mathrm{nA} \\ I_{\mathrm{inh}} &= 0.033\,\mathrm{nA} \\ \tau_{\mathrm{exc}} &= 1\,\mathrm{ms} \\ \tau_{\mathrm{inh}} &= 3\,\mathrm{ms}. \end{aligned} \tag{7.9}$$

The scaling constant A was ranging from 0.1 to 1.2 and can be thought of as the intensity of the stimulus. Further following [5], two different sets of rates were used. The first set of rates:

$$\begin{aligned} r_{\text{exc}} &= 6.88\,\text{kHz} \\ r_{\text{inh}} &= 2.88\,\text{kHz}, \end{aligned} \tag{7.10}$$

when used with $A = 1$, leads by the Campbell's theorem [34] to an average current $\pm$ standard deviation of $(0.40 \pm 0.19)\,\text{nA}$. Similarly, the second set of rates:

$$\begin{aligned} r_{\text{exc}} &= 24.52\,\text{kHz} \\ r_{\text{inh}} &= 20.52\,\text{kHz}, \end{aligned} \tag{7.11}$$

leads to the current $(0.40 \pm 0.40)\,\text{nA}$. Later in the text we will denote these methods of stimulation as factor1 and factor2.

The great advantage of using the current injection in this case is the direct connection to the experiments and parameter optimizations conducted in [5]. However, with values of A higher than 1.2 it is easy for the neuron's membrane potential to get over 0 mV. Such a high membrane potential is very unrealistic, because the reversal potential of the excitatory synapses is typically considered to be 0 mV and therefore it is impossible to hyperpolarize the membrane by such synapses above this value.

Conductance simulation

For conductance simulation we simulated the excitatory synapses as synapses with reversal potential 0 mV, peak conductance $\overline{g}_{\text{exc}} = 0.0025\,\text{mS}$ and $\tau_{\text{exc}} = 1\,\text{ms}$, inhibitory synapses with reversal potential $-80\,\text{mV}$, peak conductance $\overline{g}_{\text{inh}} = 0.0008\,\text{mS}$ and $\tau_{\text{inh}} = 3\,\text{ms}$. The values of peak conductance were chosen such that for the membrane voltage $-40\,\text{mV}$, the peak excitatory current resulting from one synaptic activation $40\,\text{mV} \cdot \overline{g}_{\text{exc}}$ would be 0.1 nA and similarly for the peak inhibitory current (from one synaptic activation) to be 0.033 nA, as was the case in the current injection experiment.

The rates used correspond to the rates in the current injection experiments. I.e. two different cases were simulated. In the first case: $r_{\text{exc}} = A \cdot 6.88\,\text{kHz}$, $r_{\text{inh}} = A \cdot 2.88\,\text{kHz}$, in the second case: $r_{\text{exc}} = A \cdot 24.52\,\text{kHz}$, $r_{\text{inh}} = A \cdot 20.52\,\text{kHz}$. A again plays the role of stimulus intensity, but ranged from 0.1 to 5.0 (we can use a much more intensive stimulus than in the case of current injection, because we don't risk running into biologically irrelevant situations).

We will denote these methods of stimulation as conductance1 and conductance2.

7.2 Subthreshold voltage integrator

In the MAT model, the membrane potential is non-resetting (as opposed to LIF), therefore it is possible to describe the whole time course of the sub-threshold membrane potential by the equation (1.15). However, as mentioned in 4.2.1, the current $I(t)$ will be simulated as a random variable and therefore it is convenient to use numerical methods.

Table 7.1: Parameters of example neurons

	α_1	α_2	ω
RS	30 mV	2.0 mV	20 mV
IB	7.5 mV	1.5 mV	19 mV
FS	10 mV	0.2 mV	10 mV
CH	−0.5 mV	0.4 mV	26 mV

The MAT model fires a spike every time the membrane potential V (2.1) reaches the value of the dynamic threshold θ (2.3). The most convenient way to check whether the threshold has been reached is to choose a sufficiently short time step Δt and check at every step whether $V \geq \theta$.

If one checks whether a threshold has been reached every time step Δt, faster convergence than $O(\Delta t)$ cannot be reached anyway. Therefore, the use of easy-to-implement-and-control numerical methods, like the explicit Euler's method (with a step size Δt), is justified.

In our simulations we will focus on the MAT* models, i.e. a specific subset of MAT where the following parameters are fixed (see page 13):

$$\begin{aligned} \tau_m &= 5\,\text{ms} \\ R &= 40\,\text{M}\Omega \\ \tau_1 &= 10\,\text{ms} \\ \tau_2 &= 200\,\text{ms} \end{aligned} \tag{7.12}$$

The free parameters of the MAT* model are α_1, α_2 and ω.

The smallest time scale of the MAT* model is $\tau_m = 5\,\text{ms}$, therefore $\Delta t = 0.1\,\text{ms}$ should be a sufficiently small time step. i_m is calculated at each step either from all active conductances, (most importantly the synaptic conductances - see eq. (1.8)) or in a simpler case is fed directly to the neuron as an input.

Throughout this chapter we will illustrate the intermediate results on four particular sets of parameters. Each set of parameters is a result of a fit performed in [5], therefore they represent real neurons in the cortex a wistar rat. The respective neurons belong to four different classes as classified by their firing activity: regular spiking (RS), intrinsic bursting (IB), fast spiking (FS), chattering (CH). Their parameters are given in the table 7.1. Note that for all the neurons the parameters describing the membrane (τ_m, R) are fixed and only the parameters of the dynamic threshold differ. Therefore, for all sets of parameters the response of the membrane will be identical.

In the figure 7.1, the response of the models with the example parameters to a step current and a stochastic input generated by the conductance simulation approach is depicted.

7.3 Conditional probability distributions

To obtain the response statistics for a given stimulus, we ran the simulation from the resting state for time $T_\text{e} + m\Delta T$, where T_e is a time interval which is long enough so that we can assume the model has reached its stationary state, ΔT is the rate coding time window (we used 500 ms) and m is the number of samples

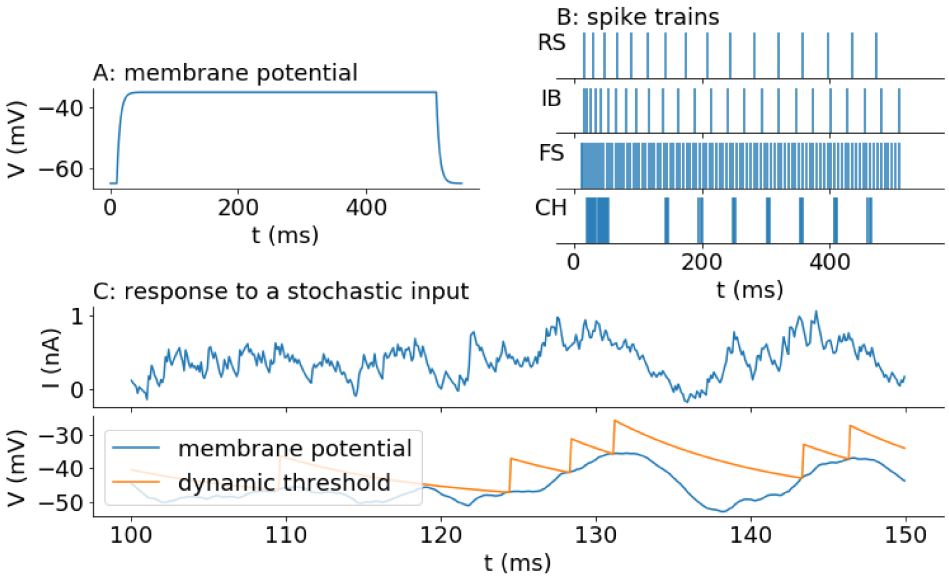

Figure 7.1: A: Response of MAT* membrane potential to a step current of 0.6 mA. B: By altering the parameters of the threshold dynamics, several characteristic spike trains may be produced as a response to the step current (RS, IB, FS, CH - description in the text). These spike trains were used to check the correctness of implementation of the MAT model, see [5, Figure 5B] C: Response of the FS model to a stochastically generated activation times of synapses, $r_{\text{exc}} = 6.88\,\text{kHz}$ and $r_{\text{inh}} = 2.88\,\text{kHz}$. The upper plot shows the time course of the total membrane current. The lower plot shows the membrane potential and the dynamic threshold. Every time the dynamic threshold reaches the potential it jumps up and a spike is fired.

we wish to obtain. The number of spikes is than computed for each of the m time windows. In our simulations we used $m = 1000$, i.e. for each considered stimulus intensity x we sampled 1000 samples of the distribution $P(n|x)$, where n is the number of spikes fired.

In all cases of input generation the scaling factor A corresponds to the intensity of the stimulus. Therefore, in this notation the conditional probability distribution we need to obtain is $P(n|A)$. In a simulation with a given stimulus type we sampled the conditional probability distribution for 100 different values of the stimulus intensity, uniformly covering the desired range (i.e. $0.1 \leq A \leq 1.2$ for the current injection simulations and $0.1 \leq A \leq 5$ for the conductance simulation runs).

In the figures 7.2 and 7.3 the conditional probability distributions for the four MAT* models defined by the table 7.1 are shown for the two different current injection regimes (factor1, factor2) and two different conducatance simulation regimes (conductance1, conductance2) defined above. The obvious difference is that there is an apparent upper limit to number of spikes fired when the conductance simulation approach is used. This is due to the fact, that the excitatory synapse potential 0 mV also sets an upper limit to the membrane potential.

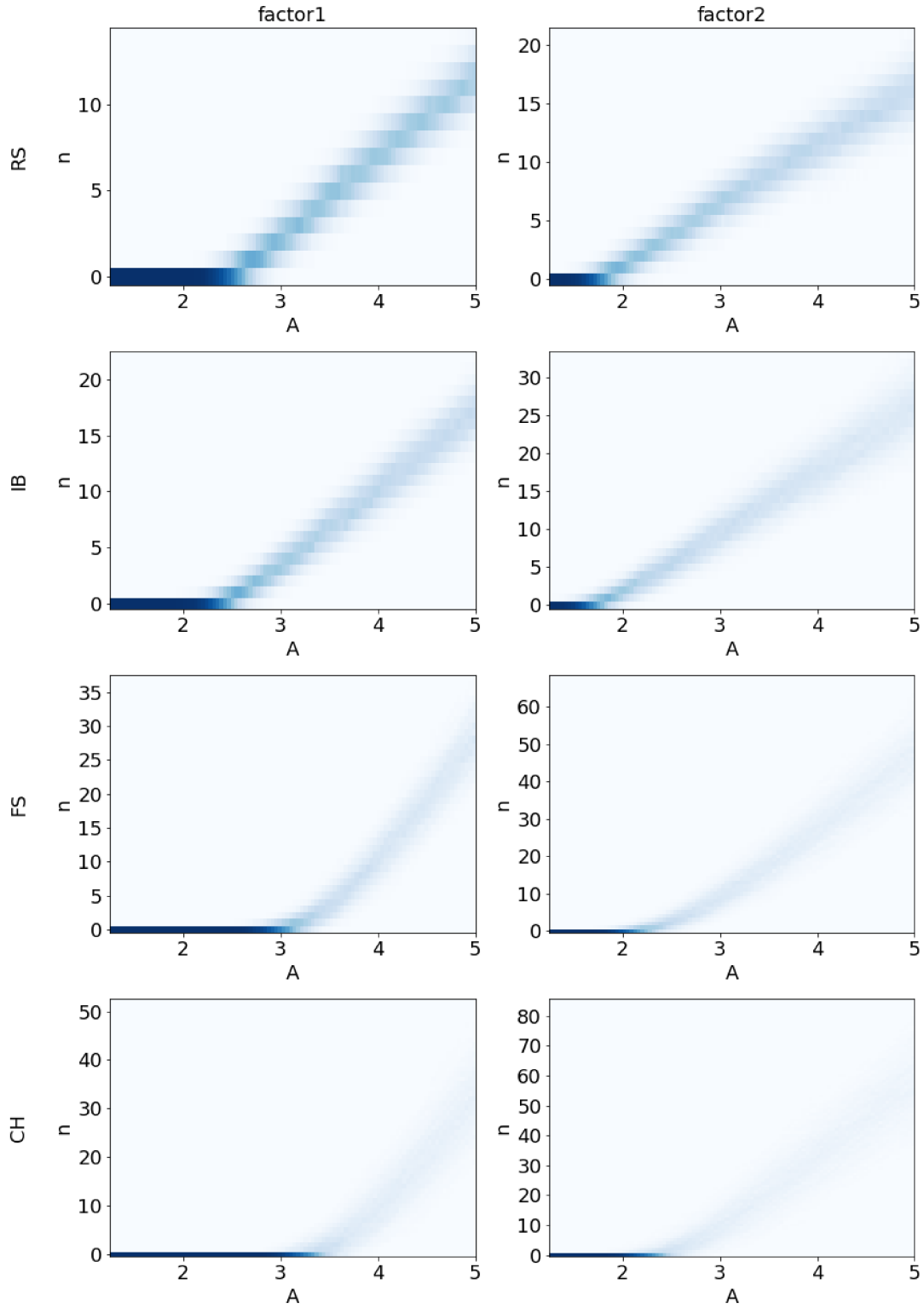

Figure 7.2: Heatmaps of the conditional probability distributions $P(n|A)$ for the input methods factor1 and factor2 for the RS, IB, FS and CH neurons defined in the table 7.1. The darker the color the higher the probability of observing given number of spikes n for a given value of stimulus intensity A.

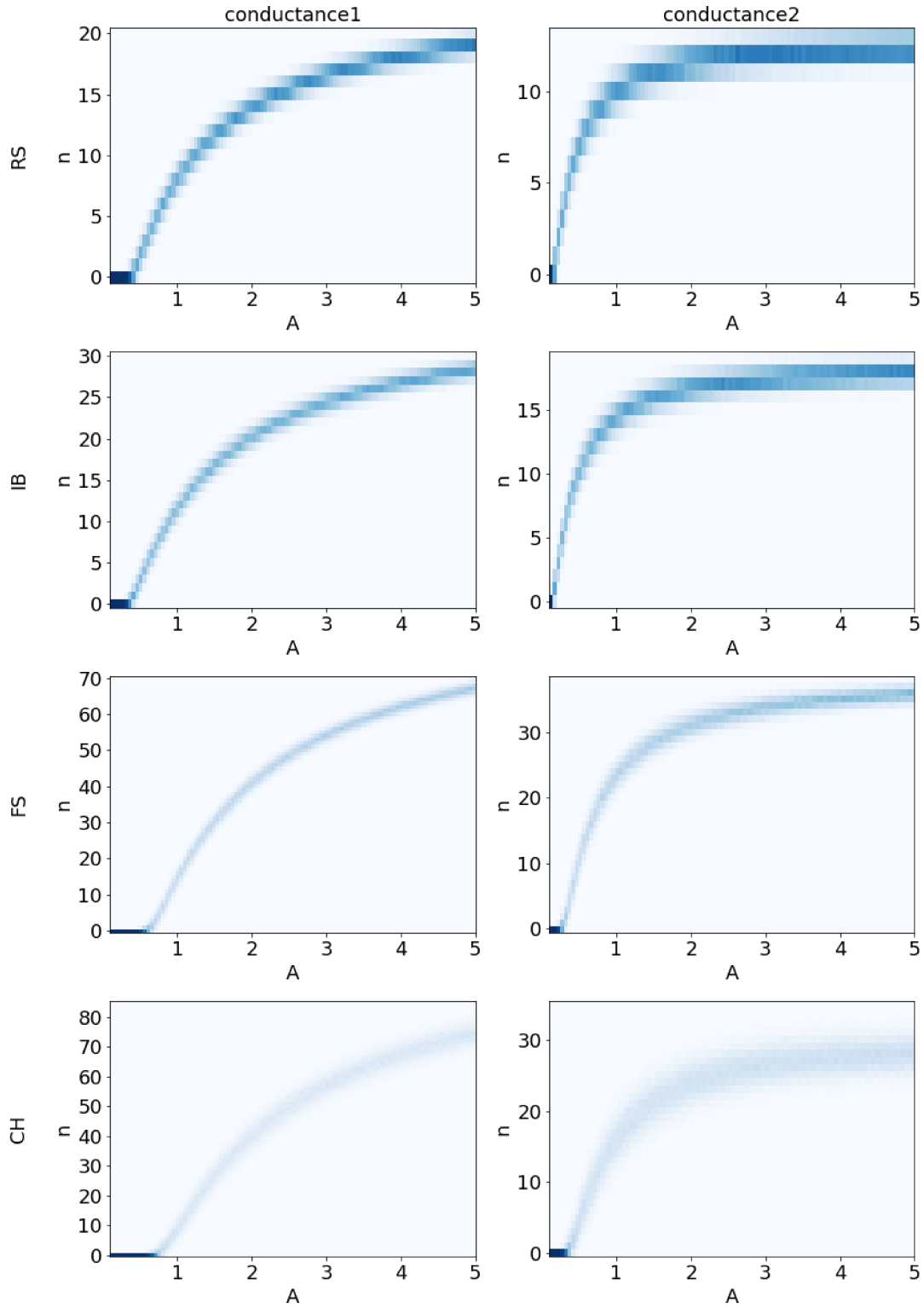

Figure 7.3: Heatmaps of the conditional probability distributions $P(n|A)$ for the input methods conductance1 and conductance2 for the RS, IB, FS and CH neurons defined in the table 7.1. The darker the color the higher the probability of observing given number of spikes n for a given value of stimulus intensity A

Table 7.2: Information capacities of the example neurons (in bits per channel use) and the respective average number of spikes fired in the optimal regime.

	conductance	conductance_2	factor	factor_2
RS	2.98 b / 9.7 sp	2.44 b / 6.2 sp	2.05 b / 5.1 sp	2.15 b / 6.9 sp
IB	3.20 b / 14.2 sp	2.63 b / 9.0 sp	2.21 b / 7.3 sp	2.26 b / 10.4 sp
FS	3.66 b / 34.3 sp	2.94 b / 18.1 sp	2.12 b / 11.5 sp	2.24 b / 18.0 sp
CH	3.00 b / 33.9 sp	2.24 b / 12.6 sp	1.74 b / 12.9 sp	1.94 b / 22.0 sp

7.4 Information capacity

We sampled the conditional probability distribution $P(n|A)$ for finitely many values of A, i.e. $\{A_1, ..., A_J\}$, therefore we were essentially working with finite input alphabet channels. However, the algorithm proposed for the continuous input alphabet channels can be still taken advantage of when modified accordingly.

We started with and an initial input alphabet $\mathcal{A}^0$ containing only the lowest and the highest value of intensity. Then at n-th step we added to the input alphabet the input symbol A_j such, that

$$A_j = \arg\max_{x \in \mathcal{A}^n} v(x; q_{p^n}). \tag{7.13}$$

I.e. we weren't searching for all local maxima, but only for one global maximum. The advantage of using this approach instead of the classical discrete Blahut-Arimoto is that from the output of the algorithm it is easier to determine which input symbols belong to the support of the optimal input distribution (which we expect to contain only a few input symbols).

From the optimization we obtain the capacity of the channel C in bits per one use of the channel. We consider a rate-coding time-window $\Delta T = 500\,\text{ms}$. If we wished to compute the capacity in bits per second, we could do so easily by computing $\frac{C}{\Delta T}$. We will, however, give the values of capacity in bits per channel use.

First, we didn't consider any constraints. In the table 7.2 the values of information capacity and the associated average number of spikes fired for the four example models and different methods of stimulation are given.

The FS neuron seems to be capable of either transmitting the most information or nearly ties break with the IB neuron, depending on the input used. However, the associated firing rates of the FS neuron in the optimal regimes are usually much higher than the rest. It is also remarkable that the capacities when the factor2 input is used (i.e. the noisier of the current injections) are higher than the capacities when factor1 input is used. However, as we will show later, this is also due to the fact that thanks to the noise the neuron is able to reach higher firing rates.

Neurons need to expand energy in order to produce spikes, therefore it can be useful to evaluate the information capacity is some constraints on the resulting firing rate are set. To take the number of fired spikes into account, we associated with each stimulus intensity the average number of fired spikes when the the neuron is stimulated with that intensity. I.e. the expense function of the neuron, same as in the case of the Poisson neuron in 4.2.1, corresponds to its tuning curve

$f(A)$:

$$e(A) = \sum_{n=0}^{+\infty} nP(n|A) = f(A). \tag{7.14}$$

We evaluated the dependence of the constrained capacity on the allowed expense by using the Blahut-Arimoto algorithm for constrained discrete channels (in combination with the proposed alphabet optimization method) for various values of the multiplier μ (evenly spaced values in the interval $[0, 1]$, 0 - which corresponds to the unconstrained problem - included). For each value of μ we obtained the expense E_μ and the constrained capacity $C(E_\mu)$ and thus we were able to interpolate $C(E)$ as a function of E.

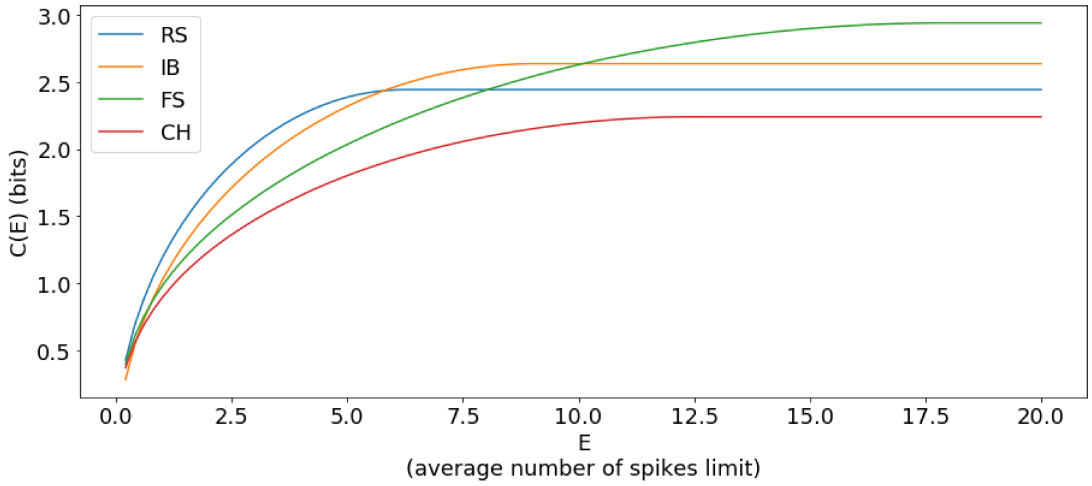

Figure 7.4: The capacity-cost function for the four example neurons given a conductance2 stimulus. When no constraints are given, the FS neuron is capable of transmitting the most information. However, as we lower the limit on average number of spikes fired, the IB and RS neurons become more effective than the FS neuron.

In the figure 7.4 the capacity-cost function is plotted for the four example neurons. Each of the neurons, except for the chattering neuron, has a specific firing rate range at which it is capable of transmitting more information than the rest. In the figure 7.5 is the capacity-cost function only of the RS neuron, but for the different stimuli. This figure suggests that the less noisy stimulus is better for information transmission (as one would expect), with the exception of the factor2 stimulus at high firing rates. This is due to the fact that the additional noise allows for higher firing rates.

In the figure 7.6 are shown the input and output distribution for the RS neuron for several different values of the constraint E. While with the current experimental methods it is not possible to measure the distribution of the inputs to the neuron, the output of neurons can be measured in vivo. If the theoretically predicted output distribution matched with the measured one, it would support the hypothesis that neurons perform optimally from information-theoretical standpoint [20], [21].

For a better feel how the parameters and the maximal allowed expense affect the capacity, we fixed the resting threshold value at $\omega = -48\,\mathrm{mV}$ and computed

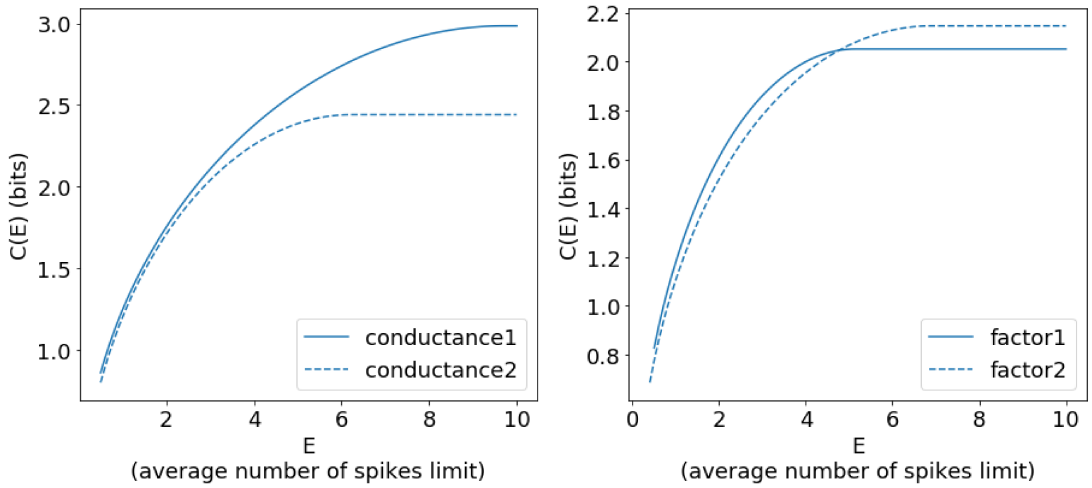

Figure 7.5: The capacity-cost function for the RS neuron - comparison of stimulation methods. The stimuli conductance2 and factor2 are generally noisier than conductance1 and factor1. This leads to worsened information transmission efficiency. However, in the case of factor2 stimulation, the noise can allow for higher firing rates and consequently allow more information to be transmitted.

the constrained capacity for a grid of (α_1, α_2) values and different maximal expenses. As an input for these simulations we used the conductance2 stimulus. We visualized the results by heatmaps shown in the figure 7.7. In the heatmaps the points corresponding to the example neurons from the table 7.1 are highlighted, however, the relation to the heatmap is only approximate, since all the neuron models have different values of the resting threshold ω.

From the heatmaps it seems, that the constraint affects greatly which parameters are optimal. For optimal information transmission at any expense any value of the parameter α_2 greater that 0 is undesirable. However, as the limit on the average expense is lowered the slow component α_2 becomes very important.

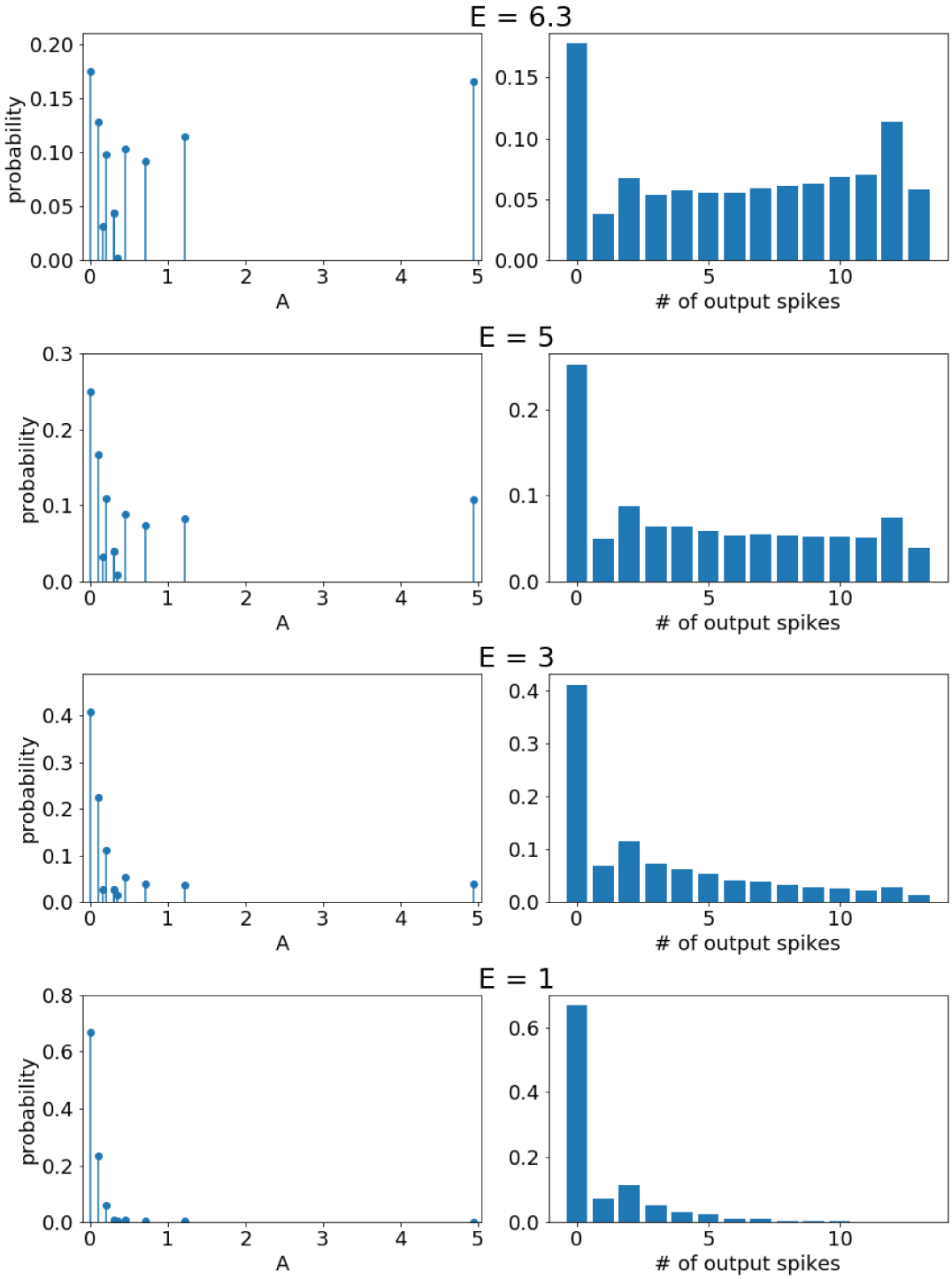

Figure 7.6: The optimal input (left columns) and output (right column) probability distribution of the RS neuron, provided the average number of output spikes does not exceed E. The topmost value $E = 6.3$ corresponds to the situation without any constrains. It can be seen that as E lowers, the probabilities of the high stimulus intensity values in the optimal regime lower as well. Comparing the distribution of outputs with experimentally measured ones can help to validate (or refute) the hypothesis, that neurons perform optimally from information-theoretical standpoint.

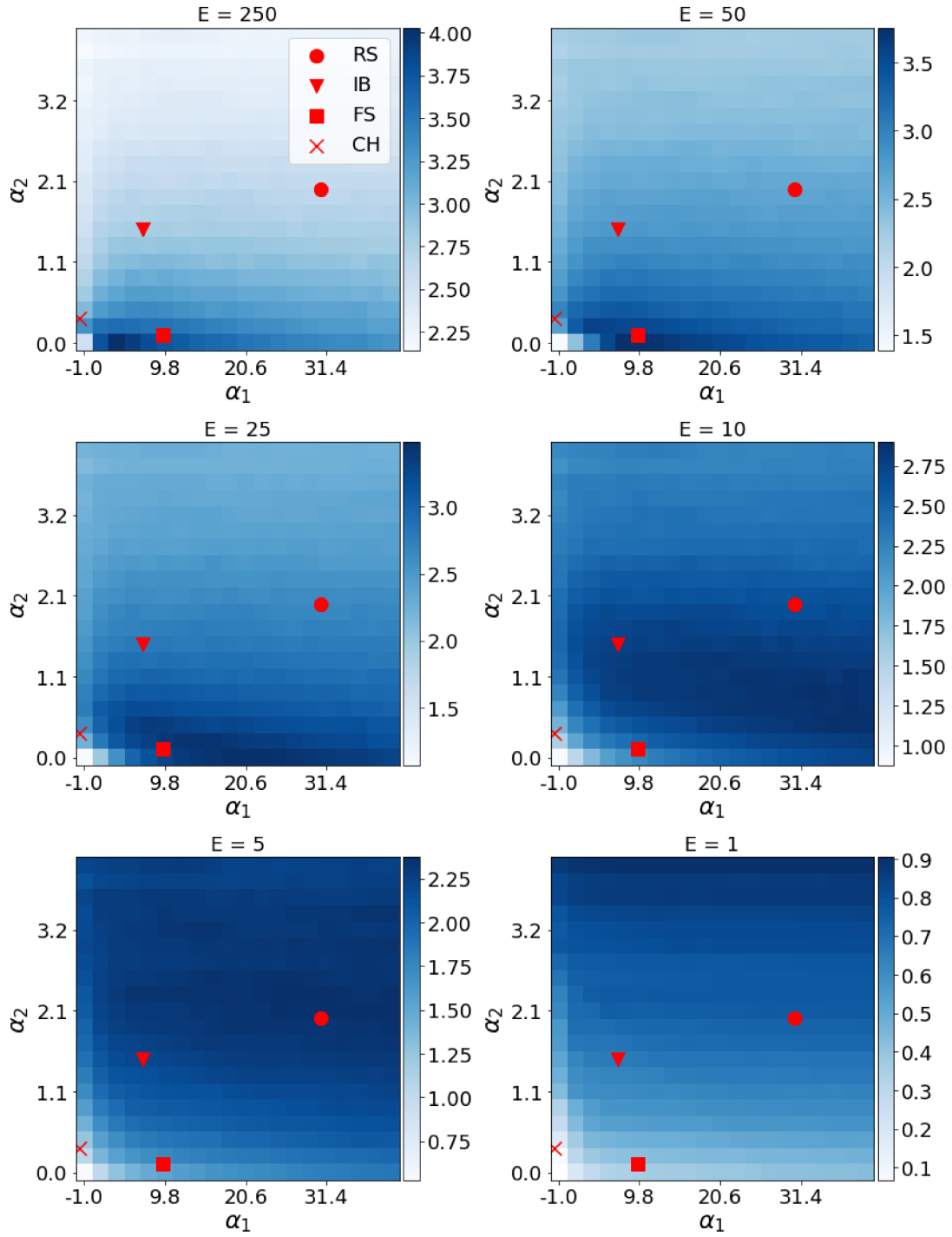

Figure 7.7: Neuronal information capacity (color, in bits per channel use) in dependence on model parameters (α_1, α_2), provided that the average number of output spikes does not exceed E. Parameter values for the four example neurons described by the table 7.1 are marked in red (RS, IB, FS, CH). The constraint $E = 250$ is equivalent to no constraint at all, since due to the refractory period, firing more than 250 spikes in 500 ms is unachievable by the model.

Conclusions

A significant portion of this work is the review of information theory and its use in neurosciences. We described the framework of Shannon's information theory with an emphasis on the converse to the noisy channel theorem. We explained what is the information capacity, first intuitively using the block coding scheme and then mathematically.

We provided an overview of mathematical properties of the mutual information both for discrete and continuous input memoryless channels, described some numerical methods of mutual information maximization and generalized an algorithm for mutual information maximization of channels with continuous input alphabet to constrained channels. We then employed this algorithm to maximize the mutual information of a Poisson neuron with a constraint on average number of spikes fired.

We then conducted Monte Carlo numerical simulations to obtain statistical properties of the relatively recently introduced MAT model [5] for a range of different parameters. The MAT model was shown to reproduce the spike times of many different neurons remarkably well [5]. This allowed us to investigate the information theoretical properties of the represented neurons. We only considered rate coding as the coding method, and we treated the neurons as discrete-time memoryless information channels without feedback.

We compared in detail four neurons, each being a representative of a different neuron class, based on their firing activity - the regular spiking (RS), intrinsic bursting (IB), fast spiking (FS) and chattering neurons. We found that each of these neurons, except for the chattering neuron, can outperform the rest if the right conditions of their average firing rates are set.

Bibliography

[1] T. J. Sejnowski, C. Koch, and P. S. Churchland, "Computational Neuroscience," Science (1988).

[2] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (The MIT Press, 2005).

[3] C. Shannon, "A mathematical theory of communication," Bell system technical journal **27** (1948).

[4] E. Lehmann and G. Casella, *Theory of Point Estimation* (Springer Verlag, 1998).

[5] R. Kobayashi, Y. Tsubo, and S. Shinomoto, "Made-to-order spiking neuron model equipped with a multi-timescale adaptive threshold," Front Comput Neurosci **3**, 9 (2009).

[6] A. F. Jahangiri and G. J. Gerling, "A multi-timescale adaptive threshold model for the SAI tactile afferent to predict response to mechanical vibration," Int IEEE EMBS Conf Neural Eng , 152 (2011).

[7] A. Borst and F. E. Theunissen, "Information theory and neural coding," Nat. Neurosci. **2**, 947 (1999).

[8] Wikimedia Commons, "A simple drawing of a neuron," (2016), file: `Neuron.svg`.

[9] R. Jolivet, R. Kobayashi, R. Rauch, R. Naud, S. Shinomoto, and W. Gerstner, "A benchmark test for a quantitative assessment of simple neuron models," J. Neurosci. Meth. **169**, 417 (2008).

[10] W. Gerstner and W. M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity* (Cambridge University Press, Cambridge, 2002).

[11] "Brian 2 example code," (2017), https://bit.ly/2ItvJrz.

[12] R. B. Stein, "A Theoretical Analysis of Neuronal Variability," Biophys. J. **5**, 173 (1965).

[13] U. Wehmeier, D. Dong, C. Koch, and D. Van Essen (MIT Press, Cambridge, MA, USA, 1989) Chap. Modeling the Mammalian Visual System, pp. 335–359.

[14] S. Yamauchi, H. Kim, and S. Shinomoto, "Elemental spiking neuron model for reproducing diverse firing patterns and predicting precise firing times," Front Comput Neurosci **5**, 42 (2011).

[15] E. D. Adrian, "The impulses produced by sensory nerve endings: Part I," J. Physiol. (Lond.) **61**, 49 (1926).

[16] E. Kandel, T. Jessell, J. Schwartz, S. Siegelbaum, and A. Hudspeth, *Principles of Neural Science, Fifth Edition*, Principles of Neural Science (McGraw-Hill Education, 2013).

[17] F. Rieke, *Spikes: Exploring the Neural Code*, A Bradford book (MIT Press, 1999).

[18] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland, "Reading a neural code," Science **252**, 1854 (1991).

[19] H. Barlow, "Possible Principles Underlying the Transformations of Sensory Messages," *Sensory Communication*, **1** (1961).

[20] J. J. Atick, "Could information theory provide an ecological theory of sensory processing?" Netw. Comput. Neural Syst. **3**, 213 (1992).

[21] L. Kostal, P. Lansky, and J.-P. Rospars, "Efficient olfactory coding in the pheromone receptor neuron of a moth," PLoS Comput. Biol. **4**, e1000053 (2008).

[22] R. G. Gallager, *Information Theory and Reliable Communication* (John Wiley & Sons, Inc., New York, NY, USA, 1968).

[23] R. Lewand, *Cryptological Mathematics*, Classroom resource materials (Mathematical Association of America, 2000).

[24] R. Ash, *Information Theory*, Dover books on advanced mathematics (Dover Publications, 1965).

[25] J. A. T. Thomas M. Cover, *Elements of Information Theory*, 2nd ed., Wiley Series in Telecommunications and Signal Processing (Wiley-Interscience, 2006).

[26] J. G. Smith, "The Information Capacity of Amplitude- and Variance-Constrained Scalar Gaussian Channels," Information and Control **18**, 203 (1971).

[27] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," IEEE Transactions on Information Theory **47**, 1290 (2001).

[28] S. Ikeda and J. H. Manton, "Capacity of a single spiking neuron channel," Neural Comput **21**, 1714 (2009).

[29] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. (John Wiley & Sons, Inc., New York, NY, USA, 1997).

[30] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, New York, NY, USA, 2004).

[31] R. Blahut, "Computation of channel capacity and rate distortion theory," *Information Theory, IEEE Transactions on*, **18**, 460 (1972).

[32] H. Witsenhausen, "Some aspects of convexity useful in information theory," IEEE Transactions on Information Theory **26**, 265 (1980).

[33] C.-I. Chang and L. D. Davisson, "On calculating the capacity of an infinite-input finite (infinite)-output channel," IEEE Transactions on Information Theory **34**, 1004 (1988).

[34] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic geometry and its applications*, Wiley series in probability and mathematical statistics: Applied probability and statistics (Wiley, 1987).