Codes and Goals of Neuronal Representations

Matthias Bethge

Universität Bremen 2003

Codes and Goals
# of Neuronal Representations

Vom Fachbereich für Physik und Elektrotechnik
der Universität Bremen

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften

— Dr. rer. nat. —

genehmigte Dissertation
von

Matthias Bethge

aus Wolfsburg

1. Gutachter: Prof. Dr. Klaus Pawelzik
2. Gutachter: Prof. Dr. Helmut Schwegler

Eingereicht am: 17.09.2003
Tag des Promotionskolloquiums: 13.10.2003

To my kids

Yara Lena and Leon Josse

Author: Matthias Bethge

Title: Codes and Goals of Neuronal Representations

Keywords: Neural code, population coding, efficient coding, population dynamics, rate coding, faithful signal transmission

PACS: 89.90.+n Other topics of general interest for physicists – 02.50.Ey Stochastic processes – 84.35.+i Neural networks 02.60.Pn Numerical optimization – 05.45.Tp Time series analysis –

Abstract: Whenever one draws conclusions from data of neuronal activity about its function for sensory processing and generating behavior, one inevitably has to hypothesize about the neural code. Moreover, most specifications of the neural code cannot be justified by data for principle reasons. Therefore, it is important to strive for descriptions of neural codes, which make their underlying assumptions as transparent as possible. This thesis combines arguments of *efficient coding* with models and constraints of *population coding* and *population dynamics* in order to derive optimal codes from a well-posed set of constraints and demands. Playing around with these assumptions uncovers their mutually dependent influence on the shape of the optimal code.

Starting from the standard model of population coding for the study of optimal tuning widths, diverging conclusions in the literature are resolved by the introduction of a new independent parameter, namely the dynamic range of a tuning function. The difficulties of applying this standard model to neuronal representations of, say natural images, motivates a more exhaustive search for characteristic features of population codes that are most relevant of coding efficiency. In fact, maximum reduction of the dynamic ranges of the tuning functions turns out to be most crucial for the maximization of Fisher information. At the same time, however, this less restricted optimization uncovers severe limitations of Fisher information as a measure for coding efficiency. In order not to rely on the heuristic argument of Fisher-optimality, direct numerical evaluations of the minimum mean square error are used (for the first time in the literature) to compare the efficiency of characteristic examples of population codes, confirming the advantage of a small dynamic range. The totality of results on optimal population coding in the first part of this thesis lead to the proposal of the *Bernoulli coding hypothesis*. In short, it states that rate coding at physiologically plausible time scales suggests the use of binary coding rather than analog coding. This statement applies to population coding as well as single neurons.

The Bernoulli coding hypothesis is challenged by criteria other than coding efficiency as well. Additionally to the study of the influence of computational constraints on the neuronal readout, the question of the robustness of a code and the possibility of *faithful signal transmission* in spite of the neuronal dynamics are investigated in the second part of the thesis. In particular the latter provides an additional strong argument for the Bernoulli coding hypothesis, which is independent from coding efficiency.

Contents

List of Figures

The strange thing is that, with so poor means, and so prosaic an imagination, we manage to formulate workable models – partially workable at least. Clever as we are at ball games and ballistics, we imagine elementary particles everywhere around us and we discuss their collisions or their capture. These naive pictures already can solve many problems.

Leon Brioullouin, 1964

Chapter 1

Introduction

1.1 Intelligence and Ecology

> When the little raven was caught by Gerkeles, the mean giant, he started to snivel: "Oh, if only I had believed my old grandma... why I haven't believed my old grandma, ...oh if only I had believed my old grandma..." The giant got sick of the cant and asked the raven: "What in God's name did your grandmother tell you?" The raven sobbed: "She said: I tell you what: once there will be a giant, and he will catch you, and then he will throw you down a kloof, and you will die overwhelmed by his laughter" and he started to snivel again: "...oh if only I had believed my old grandma..." The giant grinned maliciously. "Your Grandma shall be proved correct", he said, running to the next kloof, which was deep and steep. When he threw him down the kloof, however, the raven just opened his wings, flew a quick turn and pecked out Gerkeles eyes. Blind as he was, the giant toppled erratically over the ridge and fell down the kloof, while the raven was laughing.

Giants can be tripped up easily, when attacked in an unexpected way. It is the rule rather than the exception that their extraordinary power is bought at the cost of being too specialized, such that even slight changes in their situation suffice to make them look foolish.

Intelligence, on the contrary, rather denotes the intriguing ability to hardly ever get thrown out of the context required for successful behavior. In other words, intelligence may be defined to reflect the size of the ecological niche of an organism. Accordingly, it is an important feature of intelligence and intelligent systems to be receptive to a large number of potential interpretations for a given situation more or less simultaneously. A fascinating example for this ability is the impressive information processing performance underlying our every-day visual perception.

Egon Brunswik (1903-1955), who was a pioneer of this ecological perspective on perception [Bru34; Bru43], compared the perceiver with a boxer who is fighting to survive. Perception is highly selective, task-dependent, and speed may be more important than accuracy. As a consequence, perceiving should be treated as a series of gambles, where cues arising from the world are probabilistic in nature.

Even under highly regularized conditions, however, functional models of vision are not able to mimic human visual perception. For illustration, the task of object recognition, say to detect an animal within pictures presented on a computer screen, has not been solved yet. In contrast, human subjects are able to detect a large multitude of different objects with high speed and apparent ease.

In fact, it is difficult to avoid the conclusion that our brain 'knows' something about the structure of natural objects in general independent of the particular item to be detected. It is not trivial, however, whether it is possible to describe this knowledge in a compact way. Will we ever be able to name a rule for what makes an object an object? Or is object recognition essentially a matter of knowledge acquisition?

More concretely we can ask: is it possible to find representations of images that enhance the possibility for object recognition in general? Clearly, 'in general' can only mean 'in general for all natural objects': only if there is a specific structure exhibited by the natural environment or more precisely exhibited by those objects that are naturally relevant, is it possible to reduce the complexity of the problem.

Can we identify such structures in the set of naturally relevant stimuli? How can we explore whether the early visual system of the brain makes use of such structures?

1.2 Efficient Coding

An intriguing approach to these questions has been based on the efficient coding hypothesis, which states that sensory neurons adapt to achieve an efficient representation of sensory signals. The proposal of this idea is commonly assigned to Attneave [Att54] and Barlow [Bar59], while there is an even earlier publication by Zipf in 1949 [Zip49], which presents pioneering work in this direction, and should be mentioned, too.

The efficient coding principle allows to constrain models of sensory representations by some knowledge about the ecology of the system. One aspect of the ecology of visual perception is the fact that natural images (i.e. behaviorally relevant images) constitute a very restricted set within the space of all possible images. For illustration, one might compare the number of different images that can be generated with

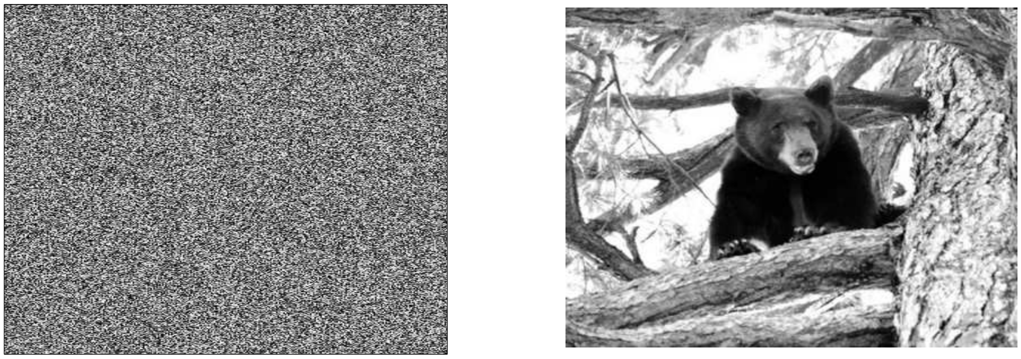

Figure 1.1. Comparison of a random grey-scale images (left) with a natural image (right). The original size of both images is identical (400 × 300 pixels).

a typical computer screen. The number of different grey-scale images that can be displayed on a computer screen is huge

$$\# \{images\} = 256^{1024\times 768} = 2^{6\,291\,456} \quad . \tag{1.1}$$

However, if you look at a random computer screen image (i.e. an image for which each pixel value is independently drawn from a uniform distribution over the 256 different possible values), it does not look very meaningful to us (see Fig. 1.1, left). In other words the great majority of images in the pixel space are not relevant, but natural images (see Fig. 1.1, right, for example) occupy a very small volume in the pixel image space only. In order to get an estimate for the number of natural images, one might assume that every natural image can be described in sufficient detail by the combination of say 10000 (English) words. Assuming further that it is sufficient to take these words out of a set of $2^{15} = 32768$ words, we end up with

$$\# \{natural\ images\} \sim 2^{150\,000} \quad . \tag{1.2}$$

Though the absolute number of natural images is still large, the information gain achieved by the constraint to have a natural image will be more than six million bits with respect to the space of arbitrary 1024 × 768 - pixel images.

In spite of the fact that the number of natural images is several orders of magnitude smaller than the number of possible images on a computer screen, it is still far too large for it to be feasible to inform a model by using the brute force method of feeding it with all natural images. Instead one can search for characteristic similarities between natural images [SLSZ03] that can be used to find more efficient image representations, for which the number of irrelevant images is reduced, to some extent at least.

One such similarity, which has been extensively studied in the recent past, is the correlation in the statistics of small patches (say 12 × 12 pixel) taken from natural

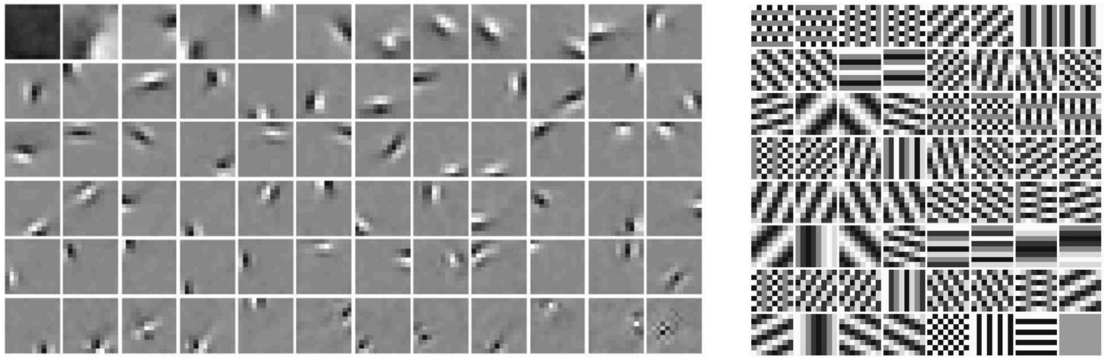

Figure 1.2. Basis functions optimized for natural image patches (left) resemble the receptive fields of cortical simple cells in contrast to the basis functions of a Fourier representation (right). Pictures are taken from a paper by Lewicki and Olshausen [LO99]

images (cf. the outstanding review paper [SO01] and references therein). The corresponding basis functions of certain linear generative models of these image patches (see Fig. 1.2) are more localized than for example a Fourier basis and resembles, surprisingly, the shape of simple cell receptive fields in cortex [OF96; OF97; BS97; vHvdS98; LO99].

1.3 Adaptive Distributed Processing and the Neural Code

The goal of this thesis is not to study the structure or the behavioral relevance of natural stimuli. Rather, the introduction should set the stage for a problem that we encounter when we seek to relate functional models of sensory processing to electrophysiological data.

In fact, there is pretty much evidence that early cortical representations are not completely predetermined by the genes, but are subjected to activity-dependent changes and reorganization [Cra99; WCF01; AH02] (see, however, also [KC02; KWGL02]). This possibility of changing has been related to Hebbian models of synaptic plasticity [MKS89; WG98] and appears to be well suited for unsupervised learning algorithms[Lin86c; Lin86b; Lin86a], which may optimize the neuronal representation according to the image statistics such that frequent patterns can be better discriminated than rare ones as predicted by the efficient coding hypothesis [KFK90; PTT00]. But is this enough to expect an agreement between the basis functions of linear generative models for patches of natural images with measured

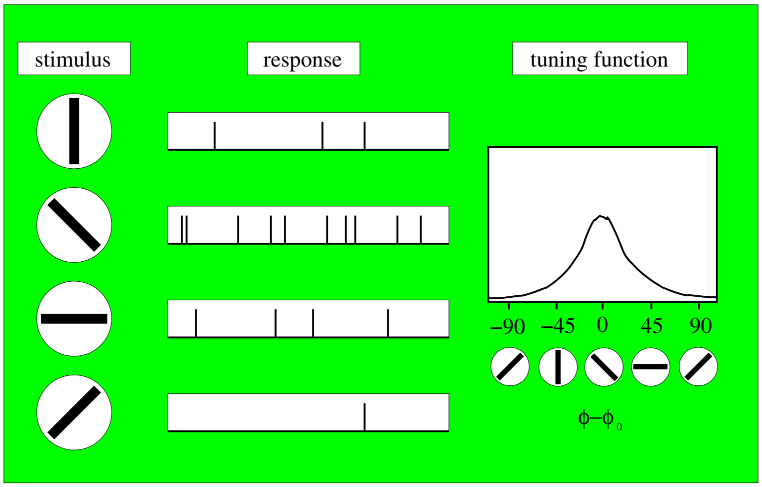

Figure 1.3. Caricature demonstrating the construction of a tuning function: the rate response of a neuron is plotted as an (interpolating) function of the stimuli used.

receptive fields in cortex?

In order to assess the meaning of such a comparison, it is necessary to understand in how far abstract generative models can be related to more detailed descriptions of neuronal responses. For example, one might ask in how far it is justified to describe neuronal rate responses by a real number. Moreover, it is not known, whether and how rate signals are relevant for neuronal processing in the cortex. In other words, if we aim to relate efficient coding models to neurophysiological data, we are faced with the fundamental problem that all models of neuronal representations rely on certain assumptions that can not be justified by data, but are used in the sense of an approximation. In fact, to date, every study[1] of cortical representations relies on such uncertain knowledge.

Generally speaking, this lack of knowledge about the fundamental question how to read the measured traces of neuronal responses is expressed by the quest for "the neural code". This quest is not so much defined by a clear posed problem, but rather constitutes a broad collection of questions arising from different contextual backgrounds.

Conceptually, this thesis developed from the background of population coding research. It contains, however, two fundamental innovations:

- A strong motivation for this work originates from the fundamental problem, how to make use of results obtained from models that suffer from a large amount of uncertainty about the assumptions used to set up the model. As a consequence it is suggested that the goal of theoretical modeling should be less restricted to the question, whether a model can reproduce an experimental

[1] It makes no difference whether the study is experimental or theoretical. Any conclusion from experimental data is possible only by the (implicit) use of a sufficiently precise model, independent of whether this model is spelled out precisely in mathematical terms, described in plain English, or not described at all.

result, but rather it should bring to light, which uncertain assumptions used by a model are most critical to reproduce the result. Most comprehensively, this way of model analysis is exemplified in the context of Fisher-optimal population codes (chapter 5).

- The other innovation is to establish a link between population coding and efficient coding models, which have been investigated to a large extent independently so far. Previous theoretical work on optimal population coding was mainly concerned with the question on how the tuning width and the noise model affect the coding accuracy of distributed neuronal representations (see chapter 4 for references and more), but ignored the influence of the shape of the signal that is encoded. Efficient coding, in turn, strongly emphasizes the effect of the particular shape of the signal that is to be encoded, but it is less detailed with respect to the neuronal constraints on signal transmission [SO01]. Since these two complementary aspects of encoding mutually influence each other, it makes sense indeed to bring together what, in fact, belongs together.

An overview of the particular problems investigated in this thesis is given in the following section.

1.4 Overview

In chapter 4, the somewhat divergent conclusions in the literature about the optimal tuning width are resolved by introducing a new relevant parameter, which is the dynamic range of the neuron's rate response. Subsequently, it is pointed out that the concept of the tuning width hardly applies to basic linear receptive field models of simple cells in visual cortex as they are commonly used in efficient coding models of image patches. In fact, the concept of *tuning functions* (see Fig. 1.3) used in population coding models has the drawback to represent a low-dimensional projection of a high-dimensional parameter space, which is determined by the selection bias of the experimentalist rather than by the search for a complete model of the behaviorally relevant parameters as it is strived for in efficient coding.

In order to account for this problem, the study of population codes is first extended to larger classes of tuning functions, where the shape of the optimal tuning functions is less predetermined by the selection of the set of candidate tuning functions (Chapter 5). Additionally, this chapter discusses the meaning of Fisher information in the context of population coding. As a result, it turns out that bell-shaped tuning functions, which are widely considered to be biologically plausible, are rather disadvantageous. Instead, tuning functions with a small dynamic range turn out to be advantageous under relevant conditions for a wide range of parameters.

The relevance of this result is further substantiated by optimizing the nonlinear gain function of a linear-nonlinear cascade neuron model, which includes the linear receptive field models used in efficient coding as a special case (Chapter 8). The crucial result of this analysis is the *Bernoulli coding hypothesis*, which says that the rate response of the individual neurons should be interpreted as binary valued ('all-or-nothing'), rather than analog. This clearly has important implications for efficient coding models and suggests a new concrete line of approach to the question, how to read neuronal rate responses.

Finally, the simplifying assumption of a discrete time, memoryless channel is relaxed and additional constraints on signal transmission due to the neuronal dynamics are investigated (chapter 11). As it turns out, the need of temporal decorrelation for rapid signal transmission imposes strong constraints on analog encoding strategies, while it is crucially less restrictive in case of Bernoulli coding.

Chapter 2

Measurement and Models of Neuronal Activity

The brain consists of a large number of coupled nerve cells (*neurons*), which for the human brain has been estimated to be of the order of 10^{10} to 10^{11} [BA91]. The most effective sort of coupling between two neurons takes place at the so called *chemical synapses*, which exclusively respond to action potentials of one of both cells only, namely the *presynaptic* neuron. *Action potentials* are characteristic electrical pulses ("*spikes*") that are actively generated by neurons, if their membrane potentials are sufficiently depolarized. Correspondingly, membrane potential fluctuations are in general divided into action potentials and subthreshold potentials, the component of the membrane potential below the *firing threshold* for spike initiation. The temporal width of spikes is about 1 ms. Although further signaling mechanisms between neurons are known, the signal transmission properties of action potentials are very distinct: While subthreshold membrane potential fluctuations can only propagate for about 1 mm before becoming substantially attenuated, signals traveling over larger distances within the nervous system are carried exclusively by action potentials. Hence, spikes are considered as the most important signal unit in neural networks of the mammalian brain.

2.1 Recording Neuronal Responses

For a direct measurement of the membrane potential, which does not only resolve the spikes, but also the subthreshold component, it is necessary to get in electrical contact with the inner part of the cell. Such *intracellular recordings* can be performed either by inserting a sharp, hollow glass electrode filled with an electrolyte into a neuron, or by attaching a broader tipped 'patch' electrode to the outer surface of the cell and breaking or perforating the membrane beneath the tip of the electrode.

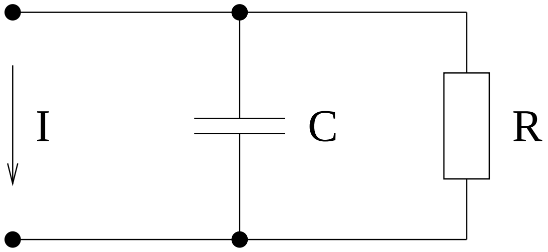

Figure 2.1. The equivalent circuit for the leaky integrate-and-fire model. The capacitance C can be charged by an external current $I(t)$, while the leak current V/R against the resistance R always discharges it.

With the patch clamp technique it is possible to either measure the membrane potential of the whole body of a cell, or of a single channel of the membrane only (cf. [SN84]).

If one is interested in *spike trains* only (i.e. the temporal sequence of action potentials), it is sufficient to bring an electrode close enough to the cell, which measures the electric field induced by the large membrane potential fluctuations of the spikes. Since such *extracellular recordings* are less complicated and more stable than intracellular recordings, they are often preferred, in particular for *in vivo* experiments. Furthermore, it is possible to record extracellularly from many neurons simultaneously, using multi-electrode arrays.

Apart from these methods, which aim to measure the membrane potentials or spikes of individual neurons, there are a few more techniques based on secondary signals that reflect certain spatial averages over the electrophysiological activity of many cells. Similar to extracellular recordings, the local field potential (LFP) and the electroencephalogram (EEG) are measurements of the electric field induced by the electric nerve cell activity, while only the size of spatial averaging is increased. Completely different is the use of voltage-sensitive optical dyes and fMRI.

2.2 Neuron Models

In order to investigate the electrical response properties of a neuron, one injects current signals into the cell during recording *in vitro*. The basic response properties of neurons find a simple description in terms of the *leaky integrate-and-fire (LIF) model* [Lap07; Tuc88; DA01; GK02]. In this model, the subthreshold membrane potential response $V(t)$ corresponds to the charging and discharging curve of a capacitance C that is wired in parallel with a resistance R (Fig. 2.1) leading to a simple low pass equation

$$\tau_{mem}\dot{V}(t) = V_{in}(t) + V_0 - V(t) \tag{2.1}$$

where the *membrane time constant* τ_{mem} is equal to the product RC and $V_{in}(t)$ denotes the input signal, which corresponds to an input current multiplied by the resistance R. Typically, values for the membrane time constant lie in the range between 10 ms and 50 ms. The resting potential V_0 is about -60 mV. Whenever the membrane potential crosses a certain threshold V_{th}, which is about 10 mV above the resting potential, the neuron generates an action potential. Within a fixed time interval $(t_{sp}, t_{sp} + \tau_{sp})$ after each threshold crossing t_{sp} the time evolution of $V(t)$ is not governed by Eq. 2.1 anymore, but given by a predefined *spike shape function* $V_{sp}(t - t_{sp})$ with $V_{sp}(0) = V_{th}$ and $V_{sp}(\tau_{sp}) = V_{reset}$. An example is displayed in Fig. 2.2, where the shape of $V_{sp} : [0, \tau_{sp}) \rightarrow \mathbb{R}$ was chosen to resemble the shape of action potentials in real measurements.

For the purpose of many theoretical studies, however, the particular shape of an action potential is not relevant. Therefore, it suffices to specify the parameters τ_{sp}, V_{th} and V_{reset} only. Then, for a constant input the time course of the membrane potential can be easily computed analytically. Let t_n, t_{n+1} denote the time instants of two subsequent spikes. Then the dynamics within $(t_n + \tau_{sp}, t_{n+1})$ is governed by the linear differential equation (2.1) with the initial condition $V(t_n + \tau_{sp}) = V_{reset}$ so that we obtain

$$V(t) - V_0 = V_{in} + (V_{reset} - V_0 - V_{in}) \exp\left\{-\frac{t - t_n - \tau_{sp}}{\tau_{mem}}\right\} \tag{2.2}$$

for all $t \in (t_n + \tau_{sp}, t_{n+1})$. Since we know that $\lim_{t \to t_{n+1}} V(t) = V_{th}$, we can now express the length $T^{ISI} = t_{n+1} - t_n$ of an *interspike interval* (*ISI*, i.e. the interval between two subsequent spikes) as a function of the constant input V_{in} and the parameters of the LIF model:

$$T^{ISI} = \tau_{sp} - \tau_{mem} \ln \frac{V_{th} - V_0 - V_{in}}{V_{reset} - V_0 - V_{in}} \tag{2.3}$$

provided $V_{in} > V_{th} - V_0$.

There are several symmetries in Eq. 2.3, which can be used to clarify the relevant determinants of the interspike interval length. By introducing new variables, $\hat{V}_{in} := V_{in}/(V_{th} - V_0)$ and $C_{reset} := 1 - (V_{reset} - V_0)/(V_{th} - V_0)$, Eq. 2.3 can be rewritten

$$\frac{T^{ISI}}{\tau_{mem}} = \frac{\tau_{sp}}{\tau_{mem}} + \ln\left(1 + \frac{C_{reset}}{\hat{V}_{in} - 1}\right) \quad , \hat{V}_{in} > 1 \tag{2.4}$$

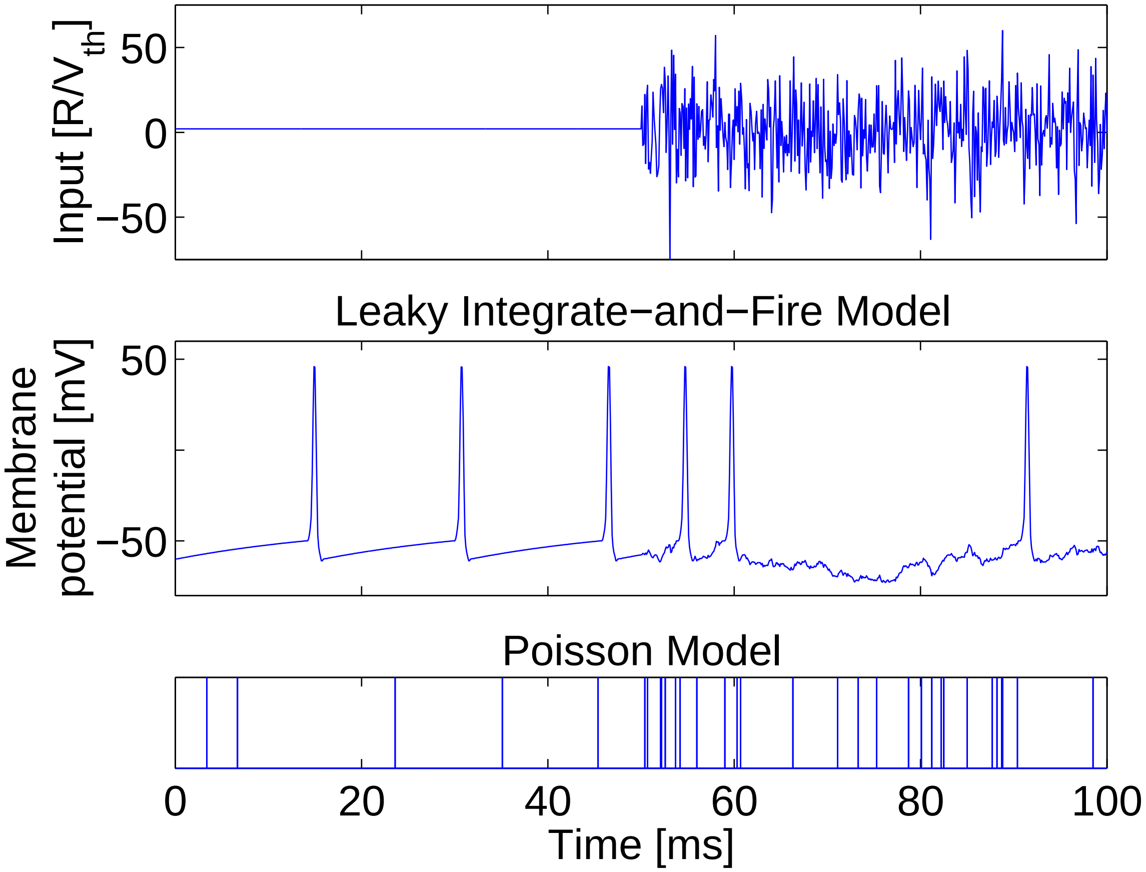

Figure 2.2. The upper panel shows the normalized input current, which equals two for the first 50 ms and then switches to a zero mean Gaussian white noise signal. The resulting membrane potential fluctuations of a LIF neuron model with $V_0 = V_{reset} = -60$ mV, $V_{th} = -50$ mV, and $\tau_{sp} = 2$ ms are shown in the middle panel. In the lower row, a random sample of a Poisson spike train is displayed, for which the rate is proportional to the input signal shown in the upper panel with the negative values treated as zero.

which means that now time is measured in units of the membrane time constant, the input is measured in units of the difference between threshold and resting potential, and the reset potential only enters Eq. 2.4 through the constant C_{reset}, which is the smaller, the larger the reset potential is.

In experimental studies it is quite common to measure and plot the frequency of spikes as a function of an injected constant input current I for characterizing the response properties of a neuron. In this way one obtains the so-called *frequency-current curve* or just *f-I curve*, which can be compared to that of neuron models Equivalently, $f := 1/T^{ISI}$ can be determined as a function of $V_{in} = RI$ for the integrate-and-fire neuron model. Since the quantities f and I are well-defined only in

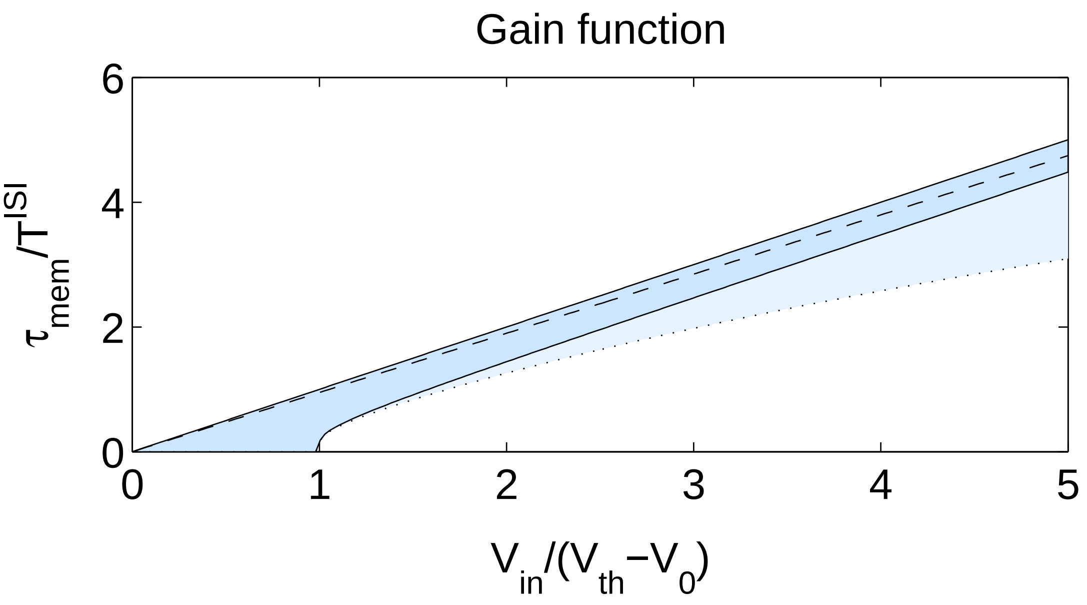

Figure 2.3. In case of $V_{reset} = V_0$ and $\tau_{sp} = 0$ the average firing rate of the LIF neuron lies within the dark shadowed region for any given (temporal) average of the input current. The dashed line indicates a tighter upper bound, which is obtained, if the maximum input $\hat{V}_{in} = 10$. The dotted lower bound corresponds to a constant current input, if $\tau_{sp}/\tau_{mem} = 0.1$.

the particular case of constant input, one may ask more generally for the relationship between the temporal averages of the rate and the input, which is called the *gain function* of the neuron. In particular, the average rate

$$\langle r \rangle = \lim_{n\to\infty} \frac{n}{\sum_{k=1}^{n} T_k^{ISI}} \quad , \tag{2.5}$$

where T_k^{ISI} denotes k-th interspike interval, is a quantity that can be measured very reliably and hence, it often makes sense to use $\langle r \rangle$ for determining the parameters of the LIF model. The relationship between $\langle r \rangle$ and the average input $\langle V_{in} \rangle$, however, is not uniquely determined by the parameters of the LIF model, but depends on the waveform of the input signal as well.

Nevertheless, it is possible to give lower and upper bounds that hold in general. Clearly, the constant current input is the least effective input, i.e. it is the waveform with the lowest average rate for a given $\langle V_{in} \rangle$. In order to obtain an upper bound on $\langle r \rangle$, we make use of the fact that in reality $V_{in}(t)$ has to be bounded so that we have $V_{min} \leq V_{in} \leq V_{max}$. The biophysical properties of neurons further suggest that the values of V_{min} and V_{max} can be related to the resting potential V_0 and the peak membrane potential $\max_{t\in[0,\tau_{sp}]} V_{sp}(t)$ that is reached during the formation of an action potential.

If we use $V_{in} \geq 0$ only, the most effective waveform is given by a Dirac delta comb $\hat{V}_{in}(t) = \sum_{\{t_{sp}\}} \delta(t - t_{sp})$, where $\{t_{sp}\}$ stands for an arbitrary set of spike times with the only restriction that $|t_{sp} - t'_{sp}| > \tau_{sp}$ for all $t_{sp}, t'_{sp} \in \{t_{sp}\}$. In this limiting case the effect of the leak current vanishes, so that the gain function of the LIF neuron becomes equivalent to that of the *perfect integrate-and-fire model* (*PIF model*), for which Eq. 2.1 is replaced by

$$\tau_{mem}\dot{V}(t) = V_{in}(t) \quad . \tag{2.6}$$

For the standard choice $\tau_{sp} = 0$ and $V_{reset} = V_0$ the gain functions of both waveforms (constant and delta comb) are shown in Fig. 2.3 (solid), enclosing a rather narrow region, within which $\langle I \rangle$ may vary for given $\langle r \rangle$, if nothing else is known about the waveform of the input current. If one makes additionally use of $\hat{V}_{in} \leq \hat{V}_{max}$, the size of this area is further reduced. Assuming $\hat{V}_{max} = 10$ according to the typical size of spikes measured *in vitro* we can compute an upper bound on $\langle r \rangle$ that is closer to the lower bound (Fig. 2.3, dashed).

It is important to note that the choice $\tau_{sp} = 0$ and $V_{reset} = V_0$, frequently taken for the sake of simplicity, have an effect on the gain function as well. While $\tau_{sp} > 0$ leads to asymptotic saturation (e.g. Fig. 2.3, dotted), $1/C_{reset}$ can be considered as a constant gain factor, because it holds

$$\frac{\tau_{mem}}{T^{ISI} - \tau_{sp}} = \left\{ \ln \left(1 + \frac{C_{reset}}{\hat{V}_{in} - 1} \right) \right\}^{-1} \approx \frac{1}{C_{reset}} \left\{ \ln \left(1 + \frac{1}{\hat{V}_{in} - 1} \right) \right\}^{-1} \quad , \hat{V}_{in} > 1 \tag{2.7}$$

In fact, the choice of C_{reset} can even change the way of signal transmission between integrate-and-fire neurons qualitatively. This qualitative change is tightly related to the possibility or impossibility of generating irregular spike trains, which has been shown to depend strongly on the reset potential [TM97]. The implications of irregular firing for signal transmission will be analyzed in chapter 11.

2.2.1 Multi-dimensional Neuron Models

Although the LIF model is motivated by the description of the membrane as a capacitance, it has to be considered rather as a phenomenological model than as a biophysical one. Historically, the *Hodgkin-Huxley* model was a great success because of its biophysical motivation [HH52]. It belongs to the class of *conductance-based models* (for a comprehensive introduction, see [DA01; GK02]), which are of the form

$$C\dot{V} = \sum_{k=1}^{n} \bar{g}_k \, m_k^{\alpha_k} \, h_k^{\beta_k} \, (V - V_k) \;\; + I(t) \quad . \tag{2.8}$$

While the membrane is modeled by a capacitance like in the integrate-and-fire neuron, the conductances governing the currents through the membrane need not be passive, but can depend on the history of the voltage time course. The different sorts of conductances commonly used can be related to certain types of ion channels in the cell membrane. Each type of ion channel can be modeled by a maximal conductance $\bar{g}_k$ and an individual reversal potential V_k. The dynamics of the conductances are described by certain gating variables m_k, h_k, for which additional differential equations have to be specified. Sometimes the gating variables are potentiated by integers $\alpha_k, \beta_k \in \mathbb{N}$, if e.g. a number of gates have to be activated simultaneously in order to open the channel. In other words, Eq. 2.8 expresses some kind of mean field approach.

Through the explicit modeling of the effective channel dynamics, it becomes possible to model the generation of a spike explicitely as the result of the autocatalytic processes caused by the voltage dependent conductances[1]. Another important aspect of conductance-based models is the fact that they can also model the ligand controlled input from the synapses more realistically, because the effect of synaptic input is not constant, but depends on the membrane potential as well. This aspect is very often neglected in network models using integrate-and-fire neurons (but see e.g. [HNT01]). Finally, there is also the possibility to account for current injection experiments, by using $I(t)$ in order to model the input from the electrode.

Different conductance-based models differ mainly in the number and types of currents taken into account. The *Steven-Conners model* e.g. has been developed as a model of the pyramidal cells in cortex, for which the Na^+ and K^+ conductances have faster kinetics than in the Hodgkin-Huxley model, which makes action potentials briefer. In addition, the Connor-Stevens model contains an extra K^+ conductance, called the *A-current*, that is transient. Further specializations and aspects, which are important for signal transmission, like *spike-frequency-adaptation (SFA)*, can be implemented in the same way by adding further currents. A comprehensive overview can be found in [DA01].

This brief report on different neuron models may already demonstrate that there is actually a large variety in the response properties of neurons, which require many specifications. Various approaches have been suggested in order to handle the large number of parameters. A well known example of a simplified conductance-based model with only two state variables is the *Fitz-Hugh-Nagumo* model [Fit61]. Later, a straightforward method to reduce the number of parameters of neuron models was published by Kepler [KAM92], which is based on time scale separation approximation. Although approximation methods are often motivated by the need to simplify

[1] The Na^+ and K^+ conductances are most important for the generation and shape of spikes

the analysis of a particular model, they typically reflect the idea that the description of some details may not be necessary, but rather obscures the actual functioning of a neuron.

2.2.2 Statistical neuron models

Statistical neuron models are prevalently used as abstract models for spike generation that aim to catch the essential dependence on some input signal of interest. While it is possible to consider a current signal injected into a neuron as such, in most experiments the input signal is defined by a parameter that distinguishes different features of externally presented stimuli. In the latter situation the neuronal response properties can be very different from those in the current injection paradigm, because they are not shaped merely by the dynamics of the individual cell anymore, but now reflect a much longer signaling pathway influenced by multiple interactions with other neuronal and synaptic processes. For this reason, a neuron responds typically quite variably to the presentation of the same stimulus signal, which makes statistical descriptions necessary.

Most statistical neuron models are build upon point process theory, for which an outstanding review paper is available [Joh96]. The most simple point process is the inhomogeneous Poisson process, which is completely characterized by an intensity function $\mu(t)$ that depends on the absolute time t only. In other words, the Poisson process is unique in its property of being completely independent from the occurrence of former spikes. An important implication of this independence on the history is the fact that for a Poisson spike train the *counting statistics*, i.e. the probability to find k spikes in any arbitrary chosen time interval (t_1, t_2), is always given by a Poisson distribution

$$P_{Poisson}(k|\lambda) = \frac{\lambda^k}{k!}\, e^{-\lambda} \tag{2.9}$$

with parameter

$$\lambda = \int_{t_1}^{t_2} \mu(t)\, dt \quad . \tag{2.10}$$

This is different for all other point processes, where the occurrence of a spike can change the shape of the intensity function $r(t|t_0, t_1, \ldots, t_n)$ with respect to t. In general, a random sample of any point process can be generated from the conditional

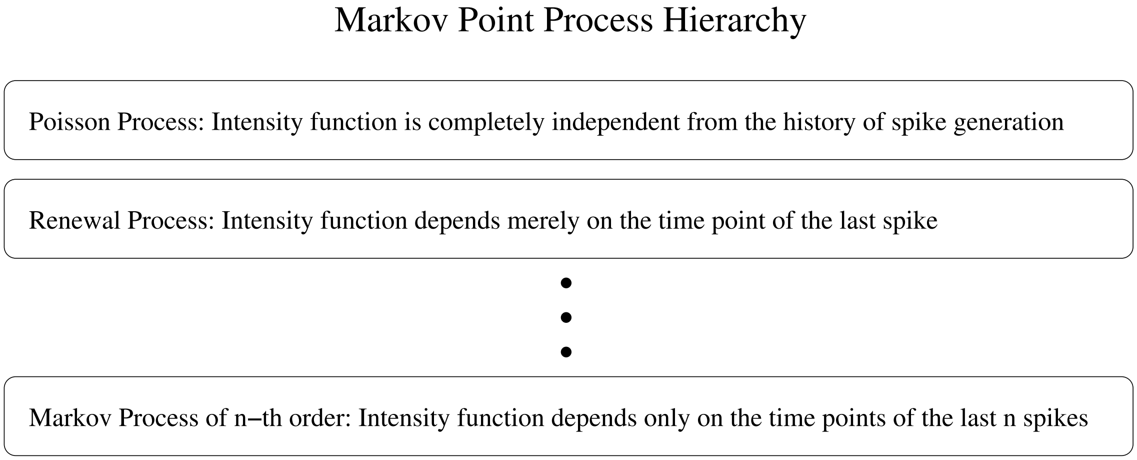
Markov Point Process Hierarchy
Poisson Process: Intensity function is completely independent from the history of spike generation
Renewal Process: Intensity function depends merely on the time point of the last spike
Markov Process of n–th order: Intensity function depends only on the time points of the last n spikes

Figure 2.4. Markov Point Process Hierarchy: Generally speaking, a point process is a Markov point process of order n, if the intensity function depends only on the last n events. The Poisson process and the Renewal process are special cases for $n = 0$ and $n = 1$, respectively.

interval distribution given by the density:

$$\rho(T|t_0, t_1, \ldots, t_n) = r(t_n + T|t_0, t_1, \ldots, t_n) \exp\left\{-\int_{t_n}^{t_n+T} r(s|t_0, t_1, \ldots, t_n)\, ds\right\} \quad . \tag{2.11}$$

Hence, the crucial question that needs to be answered in order to specify the neuron model, is how to determine the intensity function. Note, that this holds also true in case of a Poisson process, because there are many ways to define how its intensity function $\mu(t)$ shall be governed by the input. In the most simple case $\mu(t)$ is just equal to the input signal $x(t)$ for all t (which requires $x > 0$), or more generally there is a static mapping $g : \mathbb{R} \rightarrow \mathbb{R}^+$ translating the input signal into the required intensity signal (see Fig. 2.2). It is, however, also possible to model the activity of a neuron by a Poisson process, for which the intensity function $\mu(t)$ is a filtered version of the input signal. In that case spike initiation is not only governed by the current input $x(t)$, but may depend on its history in arbitrary complex ways. The choice of the Poisson model implies only that it is irrelevant whether or when the input in the history has triggered an action potential.

The basic biophysical properties of neurons, however, give reasons to make the intensity function also dependent on previous spikes. In particular, immediately after each spike there is a refractory period within which the neuron can not be driven to generate an action potential again. This piece of knowledge can be incorporated into a *renewal process.* The renewal processes constitute the next level in the hierarchy of *Markov point processes* (see Fig. 2.4), being additionally dependent on the occurrence of the last spike only. Although this additional degree of freedom is not sufficient to model prominent properties of neurons such like spike frequency adaptation exactly[2], the importance of renewal neurons lies in the fact that these mechanisms do not significantly alter the transmission of the high frequency components of the input signal.

An example of a renewal neuron model, we already know, is the LIF model introduced above, for which the period of absolute refractoriness can be modeled by the width τ_{sp} of the spike shape function and the strength of the relative refractoriness can be modeled by its reset value V_{reset}. In any case, however, the LIF model is strictly deterministic, and hence, it is an important generalization to take the possible influence of noise into account. For illustration, this can be done by defining the intensity function on the basis of LIF model via $\mu(t) = g(V(t))$ with the only difference that the spike shape function is now used exclusively to model the period of absolute refractoriness and hence, is set to V_{reset} identically. If we choose e.g. $g(V) = \beta[V - V_{th}]_+$, where

$$[x]_+ = \begin{cases} x & , \quad x > 0 \\ 0 & , \quad x \leq 0 \end{cases} \tag{2.12}$$

denotes the rectifier function, the noise level in spike generation is controlled by the parameter β, obtaining the deterministic LIF neuron as a special case for $\beta \to \infty$.

For completeness it should be mentioned that there are also point processes that are not Markovian and might be useful for neuronal modeling. For example in the Hawkes Process [Haw71] the intensity function depends on the Green's function of a linear filter responding to the Dirac impulses generated by the point process, which provides the basis for a statistical generalization of the *spike response model*, which has been used in [GvH93]. In general, the appropriate choice of a neuron model strongly depends on the question under study and has to be justified with respect to the problem at hand.

2.2.3 Large-Scale Population Models

Another line of abstraction from detailed modeling of individual cells is based on the idea, that information processing in cortex does not rely on single cells, but is based on the pooled activity of large populations of neurons. Probably the best known population model is the Wilson and Cowan model [WC72], which describes the total firing rate of a population in terms of linear response theory with a time constant typically set equal to the membrane time constant of single neurons.

[2] Exact modeling of SFA requires an explicit dependency of the intensity function on the occurrence of several spikes in the history. It is, however, possible to account for the effects of SFA in some detail just by linear or nonlinear high-pass filtering of the input[BBH$^+$01]

Population neuron models are of particular importance for large-scale neural network models, which have been extensively used to analyze pattern formation processes of cortical maps like e.g. ocular dominance [Swi80] or orientation preference maps [Swi82; Mil96; BC02]. The linear approximation of the population rate response, however, is rather weak due to the crucial dependence of the population response on the present membrane potential distribution which in general is a function of the input history [Kni72]. Recently, many theoretical studies [vS96; BH99; NB02; OKKS00; HNT01; GK02; DGM] clarified the population response properties in considerable detail using a diffusion approximation approach [GM64; Joh68]. This improved understanding also opens up new possibilities to determine relevant aspects of the dynamics of real neurons experimentally [SBM$^+$04].

The reasons for the use of population models are twofold: first, one is clearly urged to make some simplifications, if one intends to model large-scale neural network models, because otherwise such a task becomes simply intractable. On the other hand, simple or even minimalist models can have the distinct advantage to make transparent, which minimal set of assumptions are necessary in order to achieve a certain behavior of interest. In this sense, the level of abstraction of a model also tells something about what the relevant entities are. This question is indeed important for the exploration of neuronal functioning and will be the subject of the next chapter.

Chapter 3

Neural Coding

The quest for the *neural code* arises from the fact that not all known neurophysiological processes need to be relevant for the control of behavior performed by the brain. What the “right” way is to read the activity recorded from a neuron depends on the question that is to be answered. While biophysical neuron models are quite successful in reproducing the behavior of real neurons responding to injected currents, these preparations allow only very limited conclusions about the function of a cell in a (large) network of coupled neurons. In other words, the issue of the neural code constitutes an ecological approach towards the study of nerve cells, where the ’meaning’ of a neuron’s behavior is to be determined by its environment. This implies, in particular, that any sound definition of a neuronal representation requires that a model of the environment is specified, since there are typically various interesting aspects of the environment with respect to which a system can be studied.

Unfortunately, there is no established concept how to define a neural code precisely. Instead, neural coding is commonly rather introduced as a collection of particular questions (e.g. ’rate coding’ vs ’temporal coding’ or ’grandmother cells’ vs ’population coding’)[1] and methods (e.g. ’stimulus reconstruction’ experiments or ’information theory’). Since this often leads to misunderstandings between researchers, I will present my personal approach to this question. Because I am aware of the fact that any proposal for a precise grounding of this issue is at risk of being too narrow for the taste of some others, I will try to clarify its notion by the use of analogies rather than giving a formal definition.

Probably, most people will follow me if I say that the issue of neural coding is motivated by the goal of understanding how information is processed by the brain. Note that I do not claim that the brain is an information processing device *per se*. However, through the definition of a task, which defines the input (e.g. a set of visual stimuli) and the output of the brain (e.g. pressing a button or not) it is

[1] All terms will be explained below.

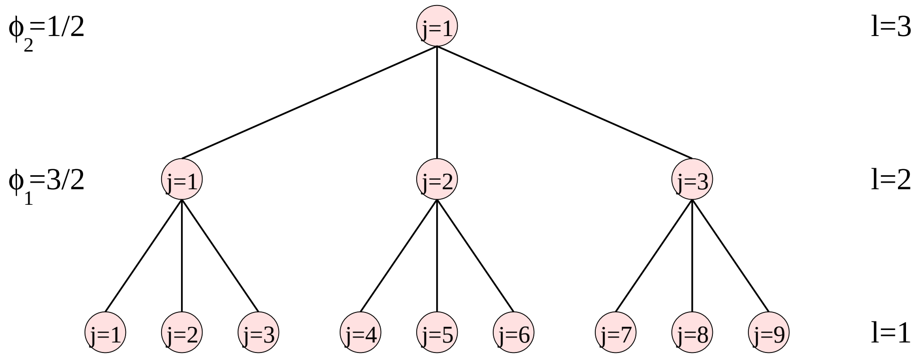

Figure 3.1. Simple feed-forward network.

often possible to describe the output as a deterministic function of the input (e.g. the button is pressed, whenever there is an animal in the presented picture). In other words, whenever we are able to prepare a system such that its response is reliably controlled by its input, we can ask for those processes in the system that are necessary and sufficient to achieve the observed input-output relationship[2].

Although artifical neural networks are too simplistic to account sufficiently for information processing in biological neural networks, we should at least be able to explain the issue of neural coding by means of a neural network model, where vague notions cannot be excused by a lack of knowledge about the system.

It is instructive to start with a very simple discrete time, feed-forward network, which is already sufficient to reveal some fundamental difficulties of defining a neural code. In this model (see Fig.3.1), each unit computes the sum of its inputs within the last time-step only and the output of the neuron is set to one, whenever this sum is super threshold, and otherwise zero:

$$s^{t+1}_{l+1,j} = \theta\left(s^t_{l,3j-2} + s^t_{l,3j-1} + s^t_{l,3j} - \phi_l\right) \quad , l \in \{1,2\} \quad . \tag{3.1}$$

In Eq. 3.1 the index l denotes the layer of the network, $j = 1, \dots 3^{(4-l)}$ enumerates the different neurons within each layer, $\phi_l = \frac{5}{2} - l$ is the threshold depending on l, and $\theta(x)$ is the Heaviside function, which is one if $x > 0$ and otherwise zero. In this way, the output of layer 3 at time t is completely determined by the activity pattern of layer 1 at time $t-2$ or equivalently by the activity pattern of layer 2 at time $t-1$. In conclusion, the functioning of this network is known, but what is the neural code of this network?

[2] Note that the notion of information processing implies that the input-output relationship is stable and hence intended to be deterministic.

3.1 Neural Code vs Neuronal Representations

Clearly, we cannot determine the relevant aspects of a neuron's activity within a network, if we had not specified before what the relevant behavior of the network is. The feed-forward structure of the network model introduced above might suggest to consider the output pattern of the third layer as the relevant network behavior. According to the point of view that a neuronal representation is something like the relevant neuronal activity used to make an intermediate result available to further processing by subsequent neurons, the binary string $(s^t_{l,1}, \ldots, s^t_{l,3^{(4-l)}})$ of neuronal activities at one particular time step in each layer would constitute an intermediate result, because this string is necessary and sufficient to determine the network output.

This is different, however, if for instance we use the network in order to control the behavior of a Khepera robot. For concreteness, let us assume the velocity v_l of the left wheel is given by

$$v^t_l = s^t_{3,1} + \sum_{\tau=0}^{100} s^{t-\tau}_{3,2} , \tag{3.2}$$

while the velocity the right wheel does not depend on the first neuron, but on the third instead:

$$v^t_r = s^t_{3,3} + \sum_{\tau=0}^{100} s^{t-\tau}_{3,2} . \tag{3.3}$$

Consequently, it is now necessary to consider the last 101 time steps $(s^{t-100}_{l,1}, \ldots, s^{t-100}_{l,3^{(4-l)}}, \ldots\ldots\ldots, s^t_{l,1}, \ldots, s^t_{l,3^{(4-l)}})$ at each layer in order to obtain a complete intermediate result. While this array of recorded activity constitutes a precisely defined *neuronal representation* of the motor behavior of the robot, it is not so much appropriate to look at if one rather seeks a simple description of the computationally relevant interaction between the individual neurons in a network. Therefore, it is worthwhile to distinguish between a *neuronal representation* and a *neural code*. While a neuronal representation should explain an external variable of interest completely, the neural code is merely about setting up an appropriate signal space.

The relation between neuronal representations and neural codes may be best illustrated by an example: Imagine an experimentalist recorded spike trains from our simple network model without knowledge about the time binning, and now tries to identify the neuronal representation of the wheel velocities on the basis of interspike intervals. Obviously, he would never be able to explore the correct neuronal representation, if he sticks to the interspike intervals as a description of the neuronal response.

In fact, every electrophysiological study relies on a hypothesis about the neural code, which makes the pursuit of the neural code such an important enterprise. As long as one is not able to set up experiments where the relationship between neuronal activity and the external variable of interest is deterministic, it is likely that at least to some extent the observed variability is due to an inappropriate choice of the signal space.

The question for the neural code is more closely related to theoretical considerations than the study of neuronal representations is. This is, because in the regime of large variability every interpretation of experimental data in terms of a certain type of neuronal representation relies on a hypothesis about the neural code that can not be justified by the data. Further clarification of this complementary relationship between neuronal representations and the neural code is possible in terms of statistical learning theory, to which we will get within the next section on neuronal decoding.

3.2 Neuronal Decoding

...The Art of Reading Thoughts

Today the prevalent experimental paradigm in neural coding research is the stimulus reconstruction approach. Instead of looking for those neuronal signals that control the behavior of the animal, the neurons are selected beforehand (by poking with an electrode into certain regions of the brain) and then it is explored which variables of a stimulus and behavior can be predicted by the recorded activity. While in principle, it would be highly desirable to correlate the neuronal activity with both, the relevant sensory signals and the behavioral signals at the same time, most studies consider only one of the two. Moreover, as a purely correlational study, stimulus reconstruction techniques can only reveal statistical dependencies, but they do not support any conclusions about causality. While the principle possibilities and limitations of this approach are illustrated in Fig. 3.2, we will now go on by taking a closer look at the inference problem that plays a major role in stimulus reconstruction studies.

Ultimately, one seeks to determine a function, which maps any possible neuronal response to the stimulus (or the motor response) that one would predict from it. Typically, the goal is to determine this function such that the predictions become as good as possible.

The issue of inference does not only arise in the context of stimulus reconstruction experiments, but it also applies for instance to the problem of perception in general, which has to be solved by the brain in order to generate successful behavior. In fact, the inferential approach constitutes a particularly important concept in understanding perceptual brain functions, because it helps to guide thinking about the problems of perception and action in general[3].

[3] A strong influence towards this way of thinking came from the studies of Helmholtz [Hel78].

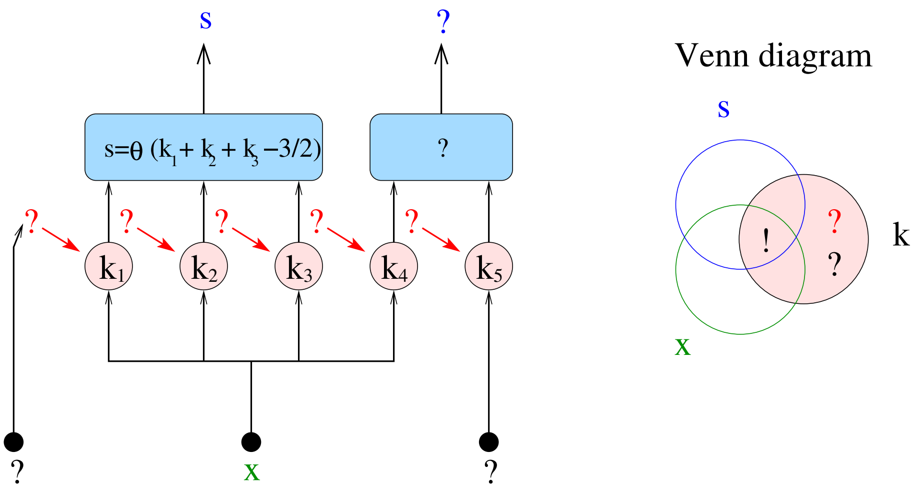

Figure 3.2. **Left**: caricature of a neural network. x denotes the stimulus and s the behavioral response controlled by the experimentalist. $k_1, \ldots, k_5$ represent examples of possibly recorded neuronal responses. To some extent, the responses $k_1, \ldots k_4$ are controlled by the stimulus, but it will be influenced by other inputs as well (red question marks). This input can be either purely endogenous or there may also be some relevant stimulus parameters that have been overlooked (black question marks). While on the basis of correlation studies k_4 can not be distinguished from, say, k_3, it is obvious in this model that k_4 is irrelevant for the considered task, while k_3 is not. Note, that "causal studies", which make use of electrical stimulation or reversible inactivation of cells, are likely going to fail to reveal a contribution of k_3, if the neuron is part of a redundant processing scheme. This case is illustrated if $k_1, k_2, k_3 \in \{0, 1\}$ are identically activated and read out as indicated. **Right**: The Venn diagram illustrates how the neuronal signals can be separated into four parts. For the signal processing only the intersection of the neuronal response k with the behavior s is relevant. This part of the signal can be further decomposed into a part, which is additionally correlated with the stimulus x, and the remaining variability, which impairs the performance of the network. The other two regions represent noise. If the noise is correlated with the stimulus, the correlation is usually called an epiphenomenon, while the remaining variability (indicated by the question marks) is not related to the considered signal processing task at all.

3.2.1 Optimal Estimation

Statistical estimation theory constitutes a successful approach to deal with inference problems. The general question in estimation theory is to determine a rule (called *estimator*) that is optimally suited to the given inference problem. Inference problems differ mainly in the kind of knowledge that is available. E.g. knowledge can vary to the extent in which the statistical relationship between observable data and the signal to be estimated is known. Therefore, the notion of an *optimal* estimator strongly depends on the assumptions one is willing to make.

The main concepts are explained in the following in a way that is intended to serve especially those readers, who want to get an overview about the main ideas in estimation theory, but do not bother too much with mathematical details. To this end, I will explain the concepts all along the same toy problem, namely to estimate the mean of a Bernoulli random variable, which constitutes a particularly simple exponential family (see appendix A.1). The most important source of the following overview is the textbook of Lehmann and Casella [LC99], which I can highly recommend for a more comprehensive and a more precise introduction.

The Bernoulli family of distributions is given by

$$P(b|f) = f^b(1-f)^{1-b} \tag{3.4}$$

where $b \in \{0,1\}$ and $f \in [0,1]$. Probably the best known estimator of the mean of a random variable is the sample mean, which in case of our example reads $\hat{f}_n = \frac{k}{n}$, where n is the total number of samples and k is the number of samples, for which $b = 1$. In the case that all samples are independent, identically distributed (i.i.d.), the sample mean is a *consistent* estimator, which means that

$$\forall \epsilon > 0 : P(|\hat{f}_n - f| > \epsilon) \overset{n\to\infty}{\longrightarrow} 0 \tag{3.5}$$

Note that the term 'estimator' is often used with two different meanings. While basically any arbitrary function of observable random variables is called an estimator, if it is intended to predict some quantity, we have here a somewhat different meaning: An estimator denoting a single function cannot be consistent, but only a particular sequence of such. Therefore, if one talks about 'the' sample mean, 'the' maximum likelihood estimator or 'the' mean square estimator, to name some examples, one does not refer to a unique function, but to a certain set of estimators, which is defined by a unique construction rule. If the latter is specified, a sequence of observations naturally leads to a unique sequence of estimators.

Like any asymptotic property, consistency is a rather weak constraint, which can be fulfilled by many functions. For illustration consider the following estimators

$$\hat{f}_n(k|a,b) = \frac{k+a}{n+b} \quad , a,b \in \mathbb{Z} \tag{3.6}$$

which all converge to the sample mean for large n and hence, are all consistent estimators as well. Loosely speaking, consistency requires that the estimator (i.e. the construction rule) yields an estimator (i.e. the function), which makes no error if noise effectively vanishes due to the large sample size.

In practice, however, one has to rather deal with situations where the amount of data is limited and the errors of the estimators are finite. In that case, it is clearly necessary to specify the loss caused by an error in order to make a comparison of two estimators possible. Therefore, the *loss function* constitutes one of the most fundamental terms in estimation theory. In general it has the form

$$l : D \times D \longrightarrow \mathbb{R} \quad , \tag{3.7}$$

if D denotes the image set of the random variable of interest. The most prominent loss function is the squared error loss $l(f, \hat{f}) = (f - \hat{f})^2$, because of its nice mathematical properties. Therefore, it will also be used in the following.

The next step is to define the *risk function*

$$r_{\hat{f}}(f) = E[l(f, \hat{f})|f] \tag{3.8}$$

in order to get a unique number representing the quality of a given estimator $\hat{f}$ at a given f (throughout the thesis $\mathrm{E}[.]$ and $\mathrm{E}[.|.]$ will be used to denoted the mean or the conditional mean of a random variable).

Spelling Eq. 3.8 out in case of the squared error loss

$$\begin{aligned}
r_{\hat{f}}(f) &= \sum_{k=0}^{n} P(k|f)(f - \hat{f}(k))^2 \\
&= \sum_{k=0}^{n} P(k|f)(f - E[\hat{f}|f] + E[\hat{f}|f] - \hat{f}(k))^2 \\
&= \sum_{k=0}^{n} P(k|f) \Big\{ (f - E[\hat{f}|f])^2 + 2(f - E[\hat{f}|f])(E[\hat{f}|f] - \hat{f}(k)) \\
&\qquad\qquad + (E[\hat{f}|f] - \hat{f}(k))^2 \Big\} \\
&= \underbrace{\Big(f - E[\hat{f}|f]\Big)^2}_{\text{“bias”}} + 2(f - E[\hat{f}|f]) \underbrace{\left(E[\hat{f}|f] - \sum_{k=0}^{n} P(k|f)\hat{f}(k) \right)}_{=0} \\
&\quad + Var[\hat{f}|f] \\
&= \mathrm{Bias}\left[\hat{f}\middle| f\right] + \mathrm{Var}\left[\hat{f}\middle| f\right]
\end{aligned} \tag{3.9}$$

yields the so-called *bias-variance decomposition* of the risk function. If the conditional mean of the estimator equals the true value of f (i.e. $E[\hat{f}|f] = f$) for all f, the bias vanishes identically and the estimator is called *unbiased*. As it is easy to check, the sample mean $f_n(k) = \frac{k}{n}$ is an unbiased estimator of f in our toy model.

In contrast to consistency, however, unbiasedness is not a minimal requirement and it is also not necessarily desirable because in general, bias and variance cannot be minimized independently of each other. This leads us to a fundamental problem,

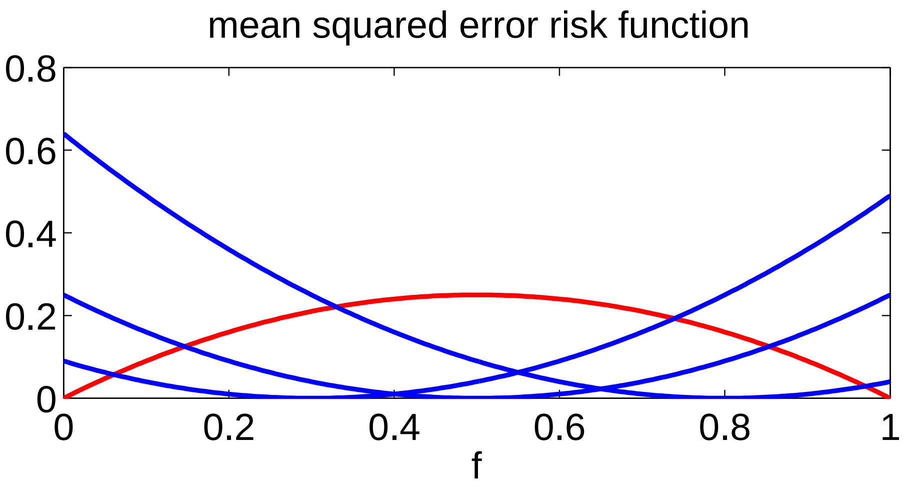

Figure 3.3. The risk functions of the sample mean for $n = 1$ (red) and of three constant estimators $\hat{f} = 0.3, 0.5, 0.8$ (blue) are compared. This illustrates that no estimator can exist, for which the risk is uniformly smaller than the risk of all other estimators. The sample mean is also the UMVU estimator and its risk is equal to the Cramer-Rao bound (red).

which can be illustrated by introducing another class of estimators, namely the constant estimators $\hat{f}_0(k) = f_0$, which do not depend on k at all. The interesting thing about such an estimator is not only that it has zero variance, but also that its risk function obviously becomes zero in case of $f = f_0$ (see Fig. 3.3). From this it follows that no estimator exists that minimizes the risk $r(f)$ uniformly for all f, apart from the trivial case, when error free estimation (i.e. $r(f) \equiv 0$) is possible.

In the special case that there is an estimator, which is **uniformly** worse than some other estimator, this estimator is called *inadmissible*. Conversely, an estimator is called *admissible*, if no estimator exists, which has a uniformly smaller risk. In other words, the derivation of a unique optimal estimator by using nothing but the risk function was possible only if one could show that all estimators but one are inadmissible. Moreover, the finding that constant estimators are always admissible as long as error-free estimation is impossible, suffices to show that further concepts are required to motivate the selection of a particular estimator. In fact, the set of admissible estimators turns out to be typically huge and thus the concept of *inadmissibility* is considered a way to rule out the most uninteresting cases before the real story starts.

In principle, there are two different approaches that enable a unique selection of an optimal estimator on the basis of the risk. Either the class of 'allowed' estimators is restricted *a priori* such that the order between the risks $r_1(f), r_2(f)$ of any two estimators becomes completely independent from f, or the estimators are compared on the basis of a loss functional $\mathcal{F}[r(f)]$ over the whole range of f. In any case, however, the unique selection of an estimator is only possible on the basis of further specifications.

One such specification, which is rather *ad hoc*, is the frequently used restriction to unbiased estimators. While unbiasedness can be motivated in order to get "impartial" estimators, it can not be seen as a **necessary** requirement of "impartiality". Nevertheless, however, it is sometimes particularly easy to determine the best one within the restricted set of unbiased estimators. Since the best unbiased estimator is clearly that one, which has the minimal variance for all f, such an estimator is called a *uniformly minimum variance unbiased* (*UMVU*) estimator. Note, however, that a UMVU estimator need not exist. E.g. no UMVU estimator exists, if one wishes to estimate $1/f$ instead of f in our toy model.

For our simple toy problem, there is a particularly elegant way to show that the sample mean even constitutes the UMVU estimator: The variance of any estimator can be bounded from below by the *Cramer-Rao bound* [AS42; Cra46; Rao46]:

$$Var[\hat{f}(k)|f] \geq \frac{(\partial_f \mathrm{E}[\hat{f}(k)|f])^2}{J[P(k|f)]} \quad , \tag{3.10}$$

where an important score function called *Fisher information* [Edg08; Edg09; Fis22] shows up in the denominator, which is defined by:

$$J[P(k|f)] \equiv E[(\partial_f \log P(k|f))^2|f] \,. \tag{3.11}$$

For unbiased estimators, the Cramer-Rao bound is completely determined by the Fisher information, because then the enumerator is equal to one. This makes Fisher information particularly useful for unbiased estimators, because its knowledge allows to bound the risk function of all unbiased estimators from below.

If f is an element of a multi-dimensional vector space, the Cramer-Rao bound can be applied to the scalar product $\langle v, f \rangle$ for any given v and hence, it is possible to use matrix algebra to efficiently write down a Cramer-Rao bound for each component with respect to a given basis set $\{v_1, \ldots, v_n\}$, which leads to the definition of the *Fisher Information matrix* [Bla88]. Since the latter is not so important for this thesis, the interested reader is refered to [Bla88; LC99] for an introduction.

For our toy problem, the Fisher information yields:

$$\begin{aligned} J[P(k|f)] &= E\left[\left(\frac{\partial}{\partial f}\left\{\log\binom{n}{k} + k\log f + (n-k)\log(1-f)\right\}\right)^2 |f\right] \\ &= E\left[\left(\frac{k}{f} - \frac{n-k}{1-f}\right)^2 |f\right] \\ &= \frac{1}{f^2(1-f)^2} E\left[(k-nf)^2 |f\right] \\ &= \frac{1}{f^2(1-f)^2} Var\left[k|f\right] = \frac{n}{f(1-f)} \end{aligned} \tag{3.12}$$

so that the risk of any unbiased estimator cannot be smaller than $f(1-f)/n$. Since the sample mean is unbiased and because its risk function equals the Cramer-Rao bound, we can conclude that the sample mean is indeed the UMVU estimator of f in our toy problem (see Fig. 3.3).

At this point it is good to place a warning, which applies to all kinds of optimization problems. It is so seductive to be satisfied with the fact that one was able to determine "the best" candidate out of a certain set that one might forget about the important aspect, how many other candidates are actually worse than the best (i.e. how large is the set with respect to which the "best"candidate has been determined). For illustration, in our example the sample mean has to be the best unbiased estimator, simply because it is in fact the only unbiased estimator. In general, the set of all unbiased estimators is rather small and can be obtained from one member of them by adding all unbiased estimators of zero (i.e. all functions $U(k)$, for which $E[U(k)|f] = 0$ for all f). In our case, the condition for an unbiased estimator of zero leads to an infinitely large set of independent linear equations

$$\sum_{k=0}^{n} \binom{n}{k} f^k (1-f)^{n-k} U(k) = 0 \quad , f \in [0,1] \, , \tag{3.13}$$

which has no solution apart from the trivial one $U(k) \equiv 0$ proving that the sample mean is the only unbiased estimator.

In fact, the restriction to unbiased estimators is quite awkward, which can best be demonstrated by a relevant example, for which it is possible to show that all unbiased estimators are so disadvantageous that they are even *inadmissible*. Recall that many situations in practice, where one wishes to estimate the parameter of a Bernoulli variable, are symmetric in the sense that one only has to consider values of f which are larger than or equal to one half. This is for instance the case, if one quantifies the performance of a subject in a binary forced choice task. If we thus consider the estimation problem with $f \in [0.5, 1]$, we have to compare the risk functions only with respect to the interval $[0.5, 1]$, and it becomes easy to find an estimator, which is uniformly better than all unbiased estimators. For illustration,

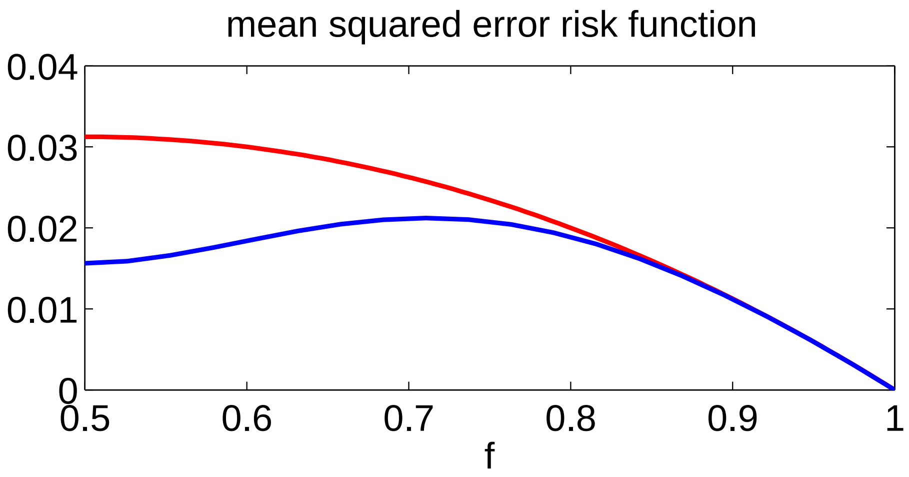

Figure 3.4. The risk function of an *ad hoc* estimator $\hat{f}(k) = 1/2 + [k/n - 1/2]_+$ (blue) is uniformly smaller than the Cramer-Rao bound (red) within the relevant interval $f \in [0.5, 1]$.

in Fig. 3.4 it is shown that the Cramer-Rao bound is uniformly larger than the risk of the biased estimator

$$\hat{f}(k) = 1/2 + [k/n - 1/2]_+ \quad . \tag{3.14}$$

where $[x]_+ = \max(0, x)$ denotes the rectifier function.

This brings us back to the question, whether it is possible to find more principled approaches to determine a good estimator. The strength of the last example lies in the fact that it demonstrates the inappropriateness of unbiasedness as an impartiality assumption without relying on any alternative definition of 'impartiality'. An alternative way, one might think of in order to determine an optimal impartial estimator, may be to use the arithmetic mean of the risk

$$\mathcal{F}_{impartial}[r(f)] = \lim_{B \to D} \frac{\int_B r(f) df}{\int_B df} \tag{3.15}$$

as objective function. In this equation $B \to D$ is a sequence of sets converging to D, which denotes the range of f. Although by now this candidate definition of impartiality is still *ad hoc*, it nevertheless sounds quite acceptable, since the risk at any possible f contributes equally to the total risk and additionally, one can show that the resulting estimators are always admissible.

At this point the educated reader of course will suspect some 'Bayesian way of thinking' in the air and I am aware of the fact that anybody who writes something about estimation theory, is still subjected to the judgment of how much 'Bayesian' or 'non-Bayesian' he or she is. My short answer to this issue is that I do not think the distinction into these two categories to be very helpful. It is more crucial to recognize instead that the interpretation of statistical models are at much higher risk to be interpreted misleadingly than deterministic models are. In particular, the common use of assumptions for the sake of mathematical tractability, simplicity, or elegance is often the source of serious pitfalls. Quite generally I would argue that the appropriateness of an assumption always depends to some extent on the entire set of assumptions that enter a model and hence, the justification of assumptions can only be evaluated with respect to the particular problem at hand. Before going deeper into the discussion what makes the difference between a strong and a weak justification of an assumption (see section (3.2.5), this section will continue first with an illustration of the Bayesian approach. To this end I resort to a somewhat artifical but nevertheless realistic example, for which it is difficult to not accept its validity.

Imagine, I programmed the incredibly exciting computer game, where you are supposed to predict the parameter f of a Bernoulli distribution from seeing n independent identically distributed samples thereof. The score you get for every single guess is proportional to $1-(f-\hat{f})^2$ and adds up linearly. Which strategy would you play, if you want to beat the high score after you have looked at the following lines of the (matlab) source code:

```
f=rand^0.5;
b(1:n)=(rand(1,n) < f);
```

Clearly, in this case it makes sense to use the fact/assumption that f is distributed according to a triangular distribution. Using the *minmax-function*

$$ {}_a[x]_b = \min(\max(a,x),b) = \begin{cases} x = a & , \quad x \leq a \\ x & , \quad a < x < b \\ x = b & , \quad x \geq b \end{cases} \tag{3.16}$$

the distribution function of the latter is given by $F(f) = {}_0[f^2]_1$, which has the density $\rho(f) = {}_0[2f]_1\theta(1-f)$. This knowledge is indeed sufficient to prove that there is a unique optimal estimator, because now by using Bayes theorem[4]

$$\rho(f|k) = \frac{P(k|f)\rho(f)}{\int P(k|f)\rho(f)df} \tag{3.17}$$

it is possible to determine the posterior risk of an estimator for any given observation:

$$r_{\hat{f}}^{post}(k) = \int r_{\hat{f}}(f)\rho(f|k)df \tag{3.18}$$

[4] The special form of the Bayes theorem used here is adapted to the toy problem, where the measure over (k,f) has a density with respect to f, but is discrete with respect to k.

The crucial advantage of the posterior risk function is that its argument is not a hidden variable as it is true for the classical risk function. Therefore, a unique optimal estimator can be defined directly as the minimizer of the posterior risk:

$$\hat{f}_{Bayes}(k) \equiv \underset{\hat{f}}{\operatorname{argmin}}\, r_{\hat{f}}^{post}(k) \tag{3.19}$$

which is called the *Bayes estimator*. As a simple consequence, the Bayes estimator is also the minimizer of the *average risk*:

$$\langle r_{\hat{f}} \rangle = \int r_{\hat{f}}(f)\rho(f)df \quad . \tag{3.20}$$

The minimum average risk is also called *Bayes risk*, because this lower bound is attained by the Bayes estimator.

In case of the mean squared error loss, Eq. 3.19 can be formally solved, which yields the *mean square estimator* (or *minimum mean square estimator*):

$$\hat{f}_{MS}(k) = E[f|k] \tag{3.21}$$

where k stands for the observation as in our example. The corresponding Bayes risk also has a quite simple form

$$\langle r_{MS} \rangle = \mathrm{E}[(f - \hat{f}_{MS}(k))^2] = \mathrm{E}[f^2] - \mathrm{E}[\hat{f}_{MS}^2(k)] \tag{3.22}$$

and is called the *minimum mean square error* (MMSE). Sometimes it is also intuitive to express the MMSE as an average over the posterior risk, which yields

$$\langle r_{MS} \rangle = E[E[(f - E[f_{MS}|k])^2|k]] = E[Var[f|k]] \, . \tag{3.23}$$

Generally speaking, Bayes estimators are difficult to compute, but sometimes solutions exist in a closed form for certain priors. Such (families of) prior distributions are called the *conjugate priors* of the given estimation problem. In our example, the conjugate prior is given by the family of *beta distributions*:

$$\rho(f|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1} \tag{3.24}$$

which includes the triangular distribution considered above as the special case with $a = 2$ and $b = 1$. For all a and b, the beta distribution has the mean

$$E[f|a,b] = \frac{a}{a+b} \tag{3.25}$$

and the variance

$$Var[f|a,b] = \frac{ab}{(a+b)^2(a+b+1)} \quad . \tag{3.26}$$

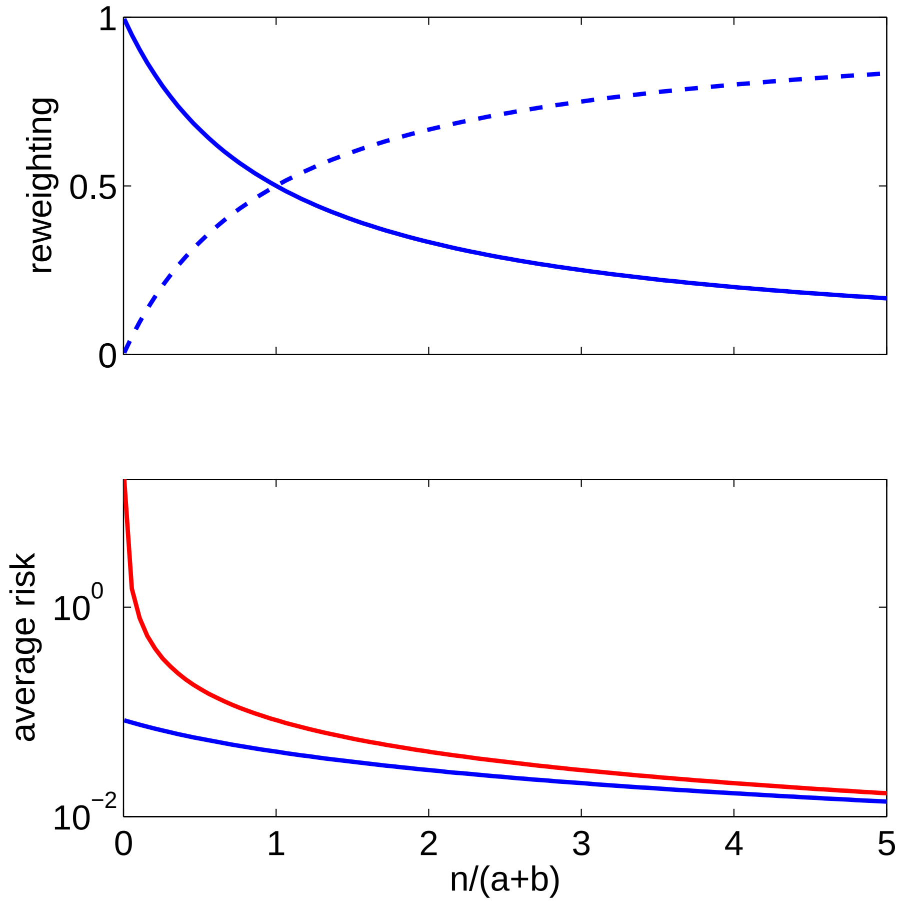

Figure 3.5. Upper panel: Course of reweighting from the constant *a priori* estimate to the sample mean as a function of $n/(a+b)$: solid line indicates the coefficient of the prior mean and dashed line indicates the coefficient of the sample mean. Lower panel: Comparison of the average risk of the Bayes estimator (blue) with the average risk of the sample mean (red) as a function of $n/(a+b)$ in case of a uniform prior (i.e. $a=b=1$).

The minimum mean square estimator in case of the beta prior reads

$$\hat{f}_{MS}(k) = \frac{k+a}{n+a+b} \quad , \tag{3.27}$$

which can be rewritten in the following decomposed way

$$\hat{f}_{MS}(k) = \frac{a+b}{n+a+b} E[f|a,b] + \frac{n}{n+a+b} \cdot \frac{k}{n} \quad . \tag{3.28}$$

This means it can be interpreted as a linear interpolation between the mean $E[f|a,b]$ of the prior distribution and the sample mean $\frac{k}{n}$, because the sum of their coefficients equals one. Particularly interesting is the behavior of the coefficients as a function of the sample size n. With increasing n the weighting moves more and more away from the prior mean closer and closer towards the sample mean. For $n = 0$ the mean square estimator is a constant exactly equal to the prior mean and in the limit $n \to \infty$ it becomes asymptotically equal to the sample mean. The exact course of the reweighting as a function of n is displayed in Fig. 3.5 together with the corresponding average $\langle r_{MS} \rangle$ over the risk of the mean square estimator

$$r_{MS}(f) = \frac{[a(1-f) - bf]^2 + nf(1-f)}{(n+a+b)^2} \tag{3.29}$$

and the average risk of the sample mean

$$\langle r_n \rangle = \frac{1}{n} \left(\mathrm{E}\left[f \middle| a, b \right] - \mathrm{Var}\left[f \middle| a, b \right] - \mathrm{E}\left[f^2 \middle| a, b \right] \right) \quad \forall n \geq 1 \,. \tag{3.30}$$

The average over Eq. 3.29 reads

$$\langle r_{MS} \rangle \;\; = \;\; \left(\frac{a+b}{n+a+b} \right)^2 \mathrm{Var}\left[f \right] + \left(\frac{n}{n+a+b} \right)^2 \langle r_n \rangle \quad . \tag{3.31}$$

which can be directly related to the decomposed representation of the MS estimator: the first term at the right hand side in Eq. 3.28 leads to a bias, but does not contribute to the error variance. The second term in turn has a positive variance, but does not contribute to the bias. Thus the reweighting between both terms of the MS-estimator directly reflects the *bias-variance trade-off* in minimum mean square estimation.

In case the Bayes estimator cannot make use of any observation, the posterior distribution equals the prior distribution and the resulting *a priori* Bayes estimator is a constant $\hat{f}_0$. The average risk of this estimator can be used as a reference in order to quantify the *loss-dependent information gain* of a measurement, which in general is given by

$$\Delta[P(k,f), l(f,\hat{f})] \equiv \inf_{\hat{f}_0} E[l(f,\hat{f}_0)] - E[l(f,\hat{f}_{Bayes}(k))] \geq 0 \quad . \tag{3.32}$$

The information gain has to be a non-negative number, because $\hat{f}_{Bayes}(k)$ constitutes the minimizer of the second term with respect to a set of functions, which contains all constant functions and hence also $\hat{f}_0$ as a special case. In case of the minimum mean squared error loss, the *a priori* Bayes estimator is simply the mean of the prior distribution so that the first term at the right hand side is equal to the variance of f. Since the second term is equal to the average over the posterior risk $E[Var[f|k]]$, the *quadratic information gain* reads

$$\begin{aligned} \Delta[P(k,f), (f-\hat{f})^2] \;\; &= \;\; Var[f] - E[Var[f|k]] && (3.33) \\ &= \;\; Var[E[f|k]] = Var[f_{MS}(k)] \quad . && (3.34) \end{aligned}$$

The consideration of the average information gain reveals an interesting fact about Bayesian estimation, namely that it separates the representation of uncertain knowledge from the problem to choose a best guess under uncertain knowledge. In other words, before the Bayesian makes a decision, he or she always determines first a probability distribution over the variable of interest. After that, the best decision is simply obtained by that estimate, which minimizes the loss under the previously determined distribution. This means that if one accepts the concept to represent uncertain knowledge about a variable by a probability distribution, it becomes an independent problem to evaluate the effective cost under a certain loss function that is due to the ambiguity represented by the distribution. This effective cost will be called the loss-dependent *uncertainty risk* in the following, which in case of the squared error loss is exactly the variance of a distribution.

In general, the *uncertainty risk* of a distribution $P(f)$ under a certain loss $l(f, \hat{f})$ is given by

$$U[P(f), l(f, \hat{f})] \equiv \inf_{\hat{f}} E[l(f, \hat{f})]_{P(f)} \quad , \tag{3.35}$$

which allows to rewrite Eq. (3.32):

$$\Delta[P(k, f), l(f, \hat{f})] = U[P(f), l(f, \hat{f})] - E[U[P(f|k), l(f, \hat{f})]]_{P(k)} \geq 0 \quad . \tag{3.36}$$

The crucial point expressed by this equation is the fact that the Bayesian framework makes sure that any increase in knowledge can never increase the average risk.

In situations like in the motivating example given above, there is actually not so much disagreement about the optimality of the Bayes estimator. The objections raised against the use of Bayesian methods rather aim at those situations in practice, where a prior distribution is not directly accessible. One aspect of the critique against Bayesian methods – maybe even the most important one – is the repeatability of the random experiment. So, while many people accept a prior distribution in case of signal transmission with a stationary source because there, one can imagine that at each instant of time a new independent, identically distributed (i.i.d.) symbol is drawn, they do not accept this approach, if the variable to be estimated is rather fixed. One reason for this distinction maybe is the fact that in case of many repetitions the empirical average over the experienced losses becomes less and less random, such that the average risk becomes a quite reliable measure for the performance of the estimator in the long run, while this is not so in case of a single experiment.

In principle, however, there is actually no difference between both situations (cf. [vNM47]). The interpretation of a time series as a one of many i.i.d. samples is never more than a Gedanken experiment. In particular the independency between subsequent observations is necessarily a purely subjective assumption of the observer. If I generate a time series, say by using the logistic map [Sch89], then to me every data point is completely dependent on the preceding one. Everybody who does not know about the origin of this time series, however, will be happy with describing it as a random variable.

Probability distributions are not given by nature, but are always the product of treating the values of a time series at different instances as unlabeled elements of the same set. Since no finite amount of data is sufficient to rule out any distribution, for which the support includes the data point, the specification of a distribution is always an extrapolation. In conclusion, in principle there is nothing new in Bayesian estimation with respect to its philosophical requirements. At most there may be a gradual change, whether the extrapolation is informed by one, a thousand or a million data points. In other words, Bayesian models should be fine, if the assumptions underlying the specification of the prior are thoroughly discussed. In fact, the major perk of Bayesian models is the enhanced visibility of the assumptions in their totality, because all assumptions apart from the likelihood and the loss function have to enter the model through the prior.

Nevertheless, it seems that the Bayesian way of reasoning gains nothing but translating a relatively simple problem (in our example selecting an element out of the unit interval $[0, 1]$) into a more difficult one, namely to select an element out of the set of all distributions over the unit interval. In other words, if we already had problems to estimate f, how can we expect to be better off in case of estimating $F(f)$?. Indeed, prior selection is not at all a simple problem and many text books circumvent a conceptual discussion of this issue, by rather turning to the style of a recipes collection and drawing too quickly on arguments of convenience.

Here, I will motivate the '*no unjustified risk reduction principle*' as I think the most satisfactory way to construct an impartial estimator, in the sense of translating the situation of having “no knowledge about the prior at all” into mathematical terms. In other words, the motivation is to find a way to construct an estimator such that this way of construction accounts for the insight that it is bad, if one has no knowledge about something, but it is even worse, if one takes a random guess as truth instead.

As discussed above, any increase in knowledge can never increase the average risk in the Bayesian framework. Conversely, this also implies that any reduction of the average risk, requires some additional amount of knowledge. Thus, if we assume that we have no knowledge about the prior at all, the average risk has to be maximal with respect to variations of the prior. In other words, the most impartial prior is the one, which does not lead to any unjustified decrease of the average risk. This leads to the definition of the *maximum Bayes risk estimator*

$$\hat{f}_{max}(k) \equiv \sup_{\Lambda} \langle r_{Bayes,\Lambda} \rangle_{\Lambda} \tag{3.37}$$

where Λ denotes any arbitrary distribution.

An important theorem (cf. [LC99]) for determining the maximum Bayes risk estimator is the following:

> *Suppose that Λ is a distribution on f such that*
>
> $$\langle r_{Bayes,\Lambda}\rangle_\Lambda = \sup_f r_{Bayes,\Lambda}(f) \tag{3.38}$$
>
> *then Λ is 'least risk reducing*[5] *and hence, the corresponding Bayes estimator is the maximum Bayes risk estimator.*

Eq. 3.38 states that the average risk of the Bayes estimator is equal to its maximum. This is the case when the risk function is constant or, more generally, when it is constant at least on a set, to which Λ assigns probability one.

Using this condition, one can show that the maximum Bayes risk estimator in our example is given by the Bayes estimator for the beta prior with $a = b = \frac{1}{2}\sqrt{n}$, because the risk function in that case (see Eq. 3.29) becomes constant. Consequently, the maximum Bayes risk estimator reads

$$\hat{f}_{max}(k) = \frac{k + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}} = \frac{1}{1+\sqrt{n}} \cdot \frac{1}{2} + \frac{\sqrt{n}}{1+\sqrt{n}} \cdot \frac{k}{n} \tag{3.39}$$

and has the risk:

$$\langle r_{max}\rangle_\Lambda = \frac{1}{4}\,\frac{1}{(1+\sqrt{n})^2} \quad . \tag{3.40}$$

If the condition of a constant risk (i.e. Eq. 3.38) holds, it implies that the maximum Bayes risk estimator is also the minimizer of the maximum risk. In fact, statisticians typically do not use the 'no unjustified risk reduction principle' to motivate the maximum Bayes risk approach, but rather start with the *minimax property*, which can be justified by the intention to minimize the risk in the worst case. A good example to motivate this goal is the problem of signal transmission with instationary sources. There one might prefer to sacrifice a little bit of the average performance in order to achieve a guarantee for the minimum quality of the signal.

Note, however, that this is a very different argument than that used to motivate the maximum Bayes risk approach. Although both approaches lead to equivalent results, whenever a constant risk estimator exists, it is important to distinguish between the different kinds of motivations. In other words, the justification of the constant risk estimator depends on the problem at hand. In the example of signal transmission with an instationary source, the minimax approach can be an appropriate, direct translation of the real problem into a mathematical model. Whenever one aims to

[5] I use 'least risk reducing' in place of the more common term 'least favorable', because the latter is associated with the concept of minimizing the worst case error, which is different from the 'no unjustified risk reduction principle'.

model the complete absence of knowledge, however, the 'no unjustified risk reduction principle' suggests to translate this situation directly into the maximum Bayes risk criterion, independent from whether it leads to the same solution as the minimax approach or not.

Nevertheless, the similarity between the maximum Bayes risk estimator and the minimax estimator suggests that the constant risk estimator has an unjustified bias towards the worst case and hence is not optimal, if one has simply no knowledge about the prior. However, if we follow the argument given above for the 'maximum Bayes risks' approach, it is not based on the worst case at all, but it only seeks to avoid risk reductions due to unjustified assumptions. The essential idea behind the 'no unjustified risk reduction principle' is that any risk reduction can only be achieved by the use of stronger assumptions and hence, the maximum Bayes risk estimator should be the least specialized one. So, could there be another reason, why the maximum Bayes risk estimator appears to be biased towards the worst case?

A textbook example to demonstrate the sub-optimality of the 'constant risk estimator' is its comparison with the sample mean in the case of our toy problem, where it turns out that the risk function of the constant risk estimator is smaller than the risk function of the sample mean within the interval $(\frac{1}{2} - c_n, \frac{1}{2} + c_n)$. Since c_n is a decreasing function of n, which is always positive, but converges to zero in the limit of large n, it seems then that the constant risk estimator becomes worse, loosely speaking, for 'almost all' f. But strictly speaking, for any finite n, there is an interval of finite length, within which the sample mean is worse. And if we do not know anything about the prior, we can also not make the assumption that the prior is flat and hence the length of the interval is indeed completely meaningless. There are uncountably many distributions which have larger mass within this small interval, in the same way as there are uncountably many distributions, which have less mass there. Without assuming a prior distribution or a hyper prior over the prior distributions (or a hyper hyper prior, etc.) there is no way to give the length of an interval (i.e. the choice of Lebesgue measure) a distinct meaning.

To make a long story short, the crucial point is that the maximum Bayes risk estimator is not impartial with respect to the choice of the loss function. For illustration, if we take the following weighted squared error loss

$$l(f, \hat{f}) = \frac{(f - \hat{f})^2}{f(1-f)} \tag{3.41}$$

which is less sensitive to errors in case of $f \approx 0.5$, the least risk reducing distribution is the uniform distribution, which is different from the 'u' shaped prior in case of the standard squared error loss.

According to the Bayesian view, however, it is natural to require that the prior and posterior distributions represent all knowledge about a variable of interest, independent of the particular choice of the loss function because as noted above, the Bayesian approach separates the optimization with respect to the loss function from the estimation of the posterior distribution. In the next section it will be explained, how the maximum Bayes risk approach can become impartial with respect to the choice of a loss function. This is often possible by applying the maximum Bayes risk approach to the estimation of the posterior distribution instead of a single valued estimate.

3.2.2 Information Theory – Part I (decoding)

Information theory can be motivated by a very Bayesian type of estimation problem. As explained above, the Bayesian always specifies first a probability distribution over the variable of interest before he or she determines the best decision on the basis of this distribution with respect to a particular loss function. In case the Bayesian is not forced to make a decision immediately, however, it makes sense to skip the second step and to postpone the decision, because in general, decisioning reduces the amount of knowledge[6].

The specification of a probability distribution itself can also be treated as an estimation problem. Instead of being forced to decide for a unique guess about the value of the variable of interest, one is allowed to give more or less ambiguous answers depending on the own degree of uncertainty. More precisely, the task in information theory is to estimate a distribution over the variable of interest rather than to estimate the value of that variable. In contrast to general estimation theory, there is a distinct loss function, to judge the quality of estimating probability distributions: This so-called *log-loss* is unique as it is the only smooth, proper, and local loss function [Ber79].

For the sake of clarity, only discrete distributions will be considered in the following. Let f denote the variable of interest and let k be the variable visible to the observer, who is required to describe his knowledge on f by specifying a (subjective) probability distribution $\hat{P}(F|k)$ for any given k. Then the log-loss is given by

$$l(f, \hat{P}(F|k)) = -\log \hat{P}(f|k) \tag{3.42}$$

where the bold font F has been used, to make clear that the observer has to estimate the entire distribution, while $P(f|k)$ with the small letter f at the right hand side denotes the probability mass of $P(F|k)$ at a particular value f.

Now one can apply the standard procedures of Bayesian estimation theory in a straightforward way. The average log risk is given by

$$\langle r_{\log} \rangle = -E[\log \hat{P}(f|k)] \quad , \tag{3.43}$$

[6] This is so, because it is not possible to combine the information of two different best guesses in a meaningful way, if nothing is known about the underlying uncertainty of these guesses.

and the posterior log risk reads

$$r_{\log}(k) = -E[\log \hat{P}(f|k)|k] \quad . \tag{3.44}$$

The uncertainty risk for the log loss is given by

$$U[P(f), -\log \hat{P}(f)] = \inf_{\hat{P}(F)} -E[\log \hat{P}(f)]_{P(f)} = -E[\log P(f)]_{P(f)} \quad , \tag{3.45}$$

which is called 'entropy' and commonly denoted by $H(F)$. Consequently, the average conditional entropy

$$H[F|K] \equiv E[H[F|k]] = E[U[P(f|k), -\log \hat{P}(f|k)]_{P(k)} \tag{3.46}$$

is nothing but the Bayes risk under log loss.

Now one can see very clearly how to deal with the prior distribution in Bayesian estimation. In case one does not know anything about the true prior distribution, the minimum loss reduction principle implies the use of the Bayes estimator, for which the Bayes risk (i.e. the average conditional entropy $H[F|K]$) takes a maximum. In conclusion, the 'no unjustified risk reduction principle' directly leads to the *maximum entropy principle* [Jay78], if applied to the problem of estimating distributions under log loss.

The derivation above does always work in case of estimation problems with a finite number of possible choices. In case of infinite problems, however, a unique and proper maximum entropy distribution only exists if additional knowledge is available. The typical way to sufficiently constrain these models is by determining certain expectation values like the variance of the distribution for instance. Well known examples of continuous maximum entropy distributions are the Normal distribution (for given mean and variance), the exponential distribution (for given mean and positivity constraint), and the uniform distribution (for a given finite support). Note, however, that not all maximum entropy distributions are well-defined. In our toy problem for instance, the uniform distribution maximizes the entropy for all distributions over the unit intervall. So one might conclude that the corresponding estimator

$$\hat{f}(k) = \frac{k+1}{n+2} \tag{3.47}$$

constitutes the most impartial choice. This conclusion, however, is not sound, because here the entropy maximization is not unique. This can be seen by recognizing the fact that one could equivalently maximize the entropy of any other parameterization, like $\tilde{f} = f^2$ for instance, which clearly allows for arbitrary different solutions.

In conclusion, in case of many estimation problems, the maximum entropy principle can provide a satisfactory solution for the construction of impartial estimators. In some cases however, like for our toy problem, the naive use of the maximum entropy method is akward, because the existence of a maximum entropy distribution itself is not enough. The method can only be applied, if the constraints governing the maximum entropy solution are specific for the considered variable.

3.2.3 Loss Functions

In this section I will just give some hints regarding the choice of the loss function. While in many cases the maximum entropy principle solves the problem to estimate the posterior distribution, independent of any loss function, the loss function is always crucial for the decision step in Bayesian estimation.

The predominant motivation for the squared error loss is in fact its mathematical convenience and frequently it does not really fit the problem at hand. While for large signal-to-noise ratios (i.e. if the average risk is close to zero) differences in the loss functions are less relevant, it matters significantly in case of large errors. Therefore, in practice one should take great care to adjust the choice of the loss function to the problem at hand. Sometimes, however, the loss function is difficult to obtain and one might wish to seek for loss functions that are more 'robust'. The concepts of 'robust loss functions', however, are by far not as satisfying as the maximum Bayes risk approach, and one is forced to use more handwaving arguments.

One such argument is the following: If one does not know anything about the loss, every error should be treated equally. Therefore, the loss function should be constant for all $\hat{f} \neq f$, which leads to the definition of the *0-1-loss*. In case f is a discrete variable, the 0-1-loss is zero if $\hat{f} = f$ and one otherwise. However, if f is a real or complex random variable, the 0-1-loss has to be modified, because otherwise the average risk would always be one for all estimators. This problem can be solved by first determining a family of optimal estimators with respect to the *ϵ-0-1-loss*

$$l_\epsilon(f, \hat{f}) = \theta(|f - \hat{f}| - \epsilon) \tag{3.48}$$

and then taking the limit $\epsilon \to 0$. In other words, it suffices if ϵ is sufficiently small that is, if all posterior distributions are sufficiently constant within any interval of length ϵ. In both cases, discrete and continuous, the resulting Bayes estimator becomes equivalent to the *maximum a posteriori estimator* (*MAP estimator*), which is very convenient and successful in practice. The MAP estimator is given by

$$\hat{f}_{MAP}(k) = \arg\sup_f \rho(f|k) \quad . \tag{3.49}$$

and contains the *maximum likelihood estimator* (*ML estimator*) as a special case, if the prior is a uniform distribution. In our toy problem the maximum likelihood estimator is equivalent to the sample mean. Note, however, that neither the MAP estimator nor the ML estimator needs to be unbiased. The most prominent counter

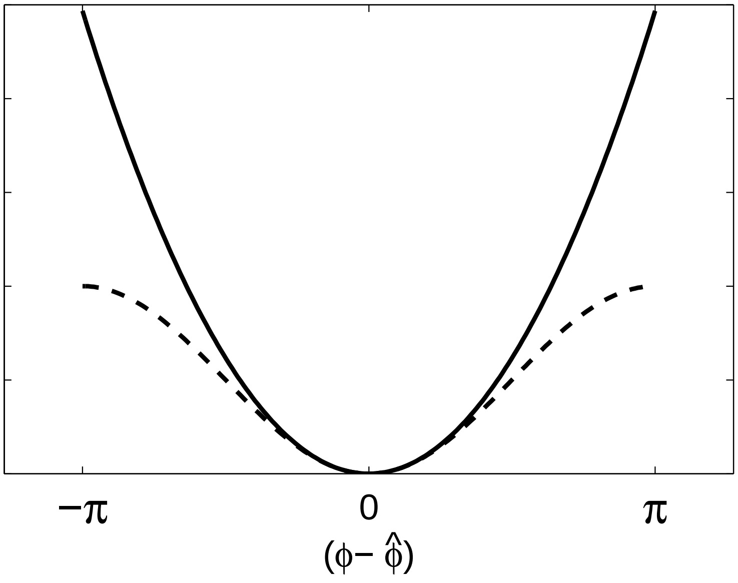

Figure 3.6.

example is the ML estimator of the variance of a Gaussian distribution, which is biased for all finite n.

Finally, it shall be mentioned that for estimation of variables living on a nonlinear manifold the loss function in the embedding space is not the same like the loss function on the lower-dimensional manifold. For illustration consider the quite frequent case of estimating a circular variable like the angle of an oriented bar stimulus, wind direction, etc. The common way to find an optimal estimator for such a variable is to optimize a two-dimensional vector estimator with respect to a loss function defined on $\mathbb{R}^2$, say the quadratic loss $l_{\mathbb{R}^2}(\vec{x}, \hat{\vec{x}}) = (\vec{x} - \hat{\vec{x}})^2$. Clearly, this is not the same as if one minimizes a quadratic loss defined directly on the manifold $\mathcal{M} := \{\vec{x} = (\cos\phi, \sin\phi)^T\}$ like $l_{\mathcal{M}}(\phi, \hat{\phi}) = (\min(|\phi - \hat{\phi}|, 2\pi - |\phi - \hat{\phi}|))^2$. Since the Euclidean distance on the unit circle between two different values of ϕ is given by

$$\left|\left|\begin{pmatrix}\cos\phi_1\\ \sin\phi_1\end{pmatrix} - \begin{pmatrix}\cos\phi_2\\ \sin\phi_2\end{pmatrix}\right|\right| = 2\sin\frac{|\phi_1 - \phi_2|}{2} \tag{3.50}$$

the effective loss function on $\mathcal{M}$ reads

$$\tilde{l}_{\mathcal{M}}(\phi, \hat{\phi}) = 2 - 2\cos(\phi - \hat{\phi}) \quad . \tag{3.51}$$

A Taylor expansion of $\tilde{l}_{\mathcal{M}}$ at $(\phi - \hat{\phi}) = 0$, however, shows that the loss function is very similar to the quadratic loss $l_{\mathcal{M}}$ for small errors

$$\tilde{l}_{\mathcal{M}}(\phi, \hat{\phi}) \approx (\phi - \hat{\phi})^2 - \frac{(\phi - \hat{\phi})^4}{12} + \mathcal{O}(6) \quad . \tag{3.52}$$

For illustration a comparison of $\tilde{l}_{\mathcal{M}}$ and $l_{\mathcal{M}}$ is plotted in Fig. 3.6

3.2.4 Statistical Learning Theory

In statistical learning theory [Vap95; SS02; Her02] the construction of an estimator itself is treated as an estimation problem. This means that it is not based on pre-known assumptions only, but can, additionally, be informed by sampling data. The framework and problems arising in statistical learning theory fit exactly the task we are faced with in neural coding. First of all, the determination of a neuronal representation means in principle to determine an estimator, which is able to predict the behavior of an animal from its neuronal activity. Stimulus reconstruction experiments also have to deal with exactly the same problem: Estimating an estimator, which predicts the stimulus as good as possible from neuronal responses. Therefore, it is worthwhile, to first understand at an abstract level, how to deal with this problem.

In contrast to general estimation theory, the special case of estimating estimators studied in statistical learning theory has a great advantage: The evaluation, how good the choice of a certain estimator was, is not so much committed to the 'right philosophy', but the quality of such a choice can be evaluated empirically by measuring the success, with which the estimator is able to predict new data.

When seeking to constrain the prediction error of a chosen estimator it becomes strikingly clear that this is only possible if one has selected a set of all candidate estimators beforehand. Without the construction of such an *hypotheses space* [Her02], it is impossible to make a reasonable choice of a particular estimator from data. The more data you are able to collect, the more candidate estimator can be compared in a meaningful way, and hence, the larger one may choose the hypothesis space. Therefore, it is a major business in statistical learning research to derive rules that tell one about the optimal complexity[7] (i.e. its 'largeness') of the hypothesis space, depending on the amount of available data. In conclusion, the success to minimize the prediction error is rather weakly influenced by data in comparison to its fundamental dependency on the pre-knowledge or *luckiness* [Her02], with which the hypotheses space has been selected.

Alluding to these complementary aspects of data and hypotheses space that together determine the best guess of an estimator in statistical learning theory, we can now better understand the relationship between the terms 'neuronal representation' and 'neural code': the 'neural code' is about the choice of the hypothesis space, while the 'neuronal representation' is exactly the function to be estimated (i.e. the function which maps the neuronal activity to the behavior of interest). Recall the example given above, where the neural code has been related to the "appropriate signal space": If the signal space is supposed to be, say, interspike intervals, this introduces

[7] In general, the hypotheses space are infinitely large sets, but similar to the state spaces in quantum mechanics, the hypotheses spaces become effectively finite, if uncountably many hypothesis can be considered to be indistinguishable.

a bias towards certain candidate functions, which, if it was carried out precisely, would correspond to the definition of a hypothesis space.

Another way to interprete statistical learning theory is that it decomposes the information underlying the outcome of an experiment into theory and data. While data have the great advantage to be irrevocable, it is often overlooked how little they are actually capable to constrain the outcome. The selection of an hypothesis space, in turn, is fundamental to the outcome but is always subjected to guessing as long as the prediction error of the estimator does not attain zero. While in the traditional fields of physics we have the satisfying situation that for several preparations the prediction error is zero (for practical purposes), this is very different in neuroscience, where the search for appropriate hypotheses spaces is still in its infancy. Clearly, the search for successful hypotheses spaces is a very difficult and arguable issue as long as 'the breakthrough' is yet to come, and I can just hope to make plausible in the next section, why I believe that the theoretical analysis of models may help to educate our guesses about neural codes.

3.2.5 Building Models from Uncertain Information

In the presentation of estimation theory, I have put strong emphasis on the conceptual issues regarding the question, how to translate a given situation into a statistical model. In this section, the issue of model selection will be discussed at a more general level with a particular focus on the problem of dealing with imprecise or uncertain knowledge. Building models from uncertain information always bears some difficulties and potential disagreement between researchers. This problem is not specific to estimation theory, but arises whenever we want to set up precise models out of imprecise information, and in fact there is hardly any agreement in neuroscience about how to evaluate the appropriateness of neural network models.

At this point, one might like to ask whether it makes sense to set up precise (mathematical) models at all, if too many ingredients are necessarily subjected to guessing. In particular, experimentalists tend to have doubts as to the use of precise models in neuroscience, because they feel that too many assumptions lack experimental justification. Indeed, it is true that the gap between theory and experiments is so large that – to be honest – critical experimental testing of electrophysiological predictions of a model is nothing but a graceful wish today, whenever these predictions go beyond current injection studies at isolated cells. A clear barrier for example is set by the impossibility to monitor the synaptic inputs of a cell caused by the presynaptic activity in the network within which the cell is embedded.

Due to this lack of critical testing in neuroscience, we have the dissatisfying situation that models do not compete for the explanation of data, but there is enough place for all models that somehow fit to the data of interest. In this way, models (as well as vague proposals) accumulate more and more, which often obscures the available information rather than making it more transparent and instructive.

So why bother with precise models and the resulting mathematical difficulties? What can be learned from models, for which almost all assumptions are wrong or which are at best a caricature of the true situation? A simple reason to begin with is that specifying precise models is healthy in the sense that it reveals which kind of assumptions we typically make without recognizing it. Indeed, one can frequently observe the beliefs, that a model could rely on a smaller or larger amount of assumptions[8]. I even saw, how models have been advertised to be completely "free of assumptions".

The deeper reason, why I indeed believe that the study of precise models is worth bothering with, may be illustrated as follows: Like in the case of optimal estimation, the effect of an assumption relies on the totality of specifications and can not be evaluated without a precise model. Some assumptions will turn out to be very critical, while some other assumptions can be changed influencing the outcome of the model only weakly. The question, whether the use of an approximation in case of uncertain knowledge can be justified is not a matter of the amount of uncertainty in the first place, but can be evaluated only if one knows about the sensitivity with which the result depends on this assumption. In other words, it is the uncertainty of the result caused by an uncertain assumption of the model, which tells whether its justification is strong or weak. Therefore, the understanding, which of the assumptions are most critical constitutes an important guide for further research, for it tells something about which sort of data may be most informative to the model.

A colleague and good friend asked me once: “Have you ever seen real data?” I responded: “Well,... yes, but I have to admit, it could have been more. On the other hand, let me also ask back: Have you ever checked thoroughly what you can conclude from your measurement by simulating a poke of your electrode into a model network first?”

This is the spirit that was governing my work presented in this thesis. Starting from some kind of standard models, I tried to figure out how far the results are affected by variations in the individual assumptions. Gambling a little bit around with assumptions can tell us how critically they influence the results. Replace the system of interest by a caricature model and figure out whether the terms commonly used make sense at least in the toy model. Indeed, it can happen that researchers try to support their favorite ideas about the neural code by setting up a strawman (an uninteresting null hypothesis) that sounds like the opposite of the preferred hypothesis, but is actually much more specialized. For example, though nobody believes that neurons are Poisson process generators, a lot of work has been done to

[8] Clearly, there is no absolute measure for the amount of assumptions in a model, but the latter can be measured only with respect to a particular class of models that itself has to be specified beforehand.

falsify this idea in order to suggest that there is some kind of 'temporal coding' at work. In fact, however, it does not reveal any interesting kind of 'synchronization' or 'temporal coding', but only what we actually know, namely that neurons are no Poisson process generators.

I hope that the investigations and discussions in this thesis will encourage other researchers, to report more explicitely on the degree of uncertainty of the individual ingredients that entered their models[9]. It does not seem to be sufficient, if everybody is left to think about the choice of assumptions himself, but I believe that it is important that we also write and talk about it. In particular, assumptions that are weakly justified and potentially critical for the results should be pointed out and discussed. While it is understandable that there is a tendency to rather hide such 'problematic' assumptions, it should actually be appreciated as a significant scientific effort that deserves to be acknowledged.

3.3 Systematic Description of Neural Codes

This section provides a short overview of a set of questions that may be suitable to decompose the pursuit for the neural code into a set of sub-problems. Recall that I defined a neural code as a specification of the signal space describing how to read or represent the neuronal activity. A prominent way to do this is the *peri stimulus time histogram (PSTH)*, where time is discretized by slicing it into bins of equal width and counting the number of spikes within each time bin. In this way, one obtains a vector of integer valued numbers, which could be called a "PSTH code". Note, however, that there are critical parameters to be specified. How large is the width of a bin? How many bins are taken (i.e. how large is the dimension of the resulting signal vector)? How to define the latency (i.e. the relative time delay between stimulus signal and neuronal PSTH response)? These and similar question are discussed in the following.

3.3.1 Spike Timing Precision

The question for the relevant spike timing precision seems to be a good one to start with. A simple way to implement an assumption about the required temporal resolution, is to discretize time by slicing it into bins of equal width, exactly as in case of the PSTH. The higher the required temporal resolution, the smaller the bin width to be chosen. This choice of the 'sampling frequency' is a fundamental requirement and has to be specified for 'rate codes' as well as for any other coding

[9] Even if the model is not spelled out in mathematical terms, there is also a (vaguely specified) model, whenever one draws conclusion from an experiment.

scheme. Therefore, a systematic description of a neural code should always start with specifying the temporal precision. In other words, every (reasonable) code we can think of can be expressed as a function of a time histogram over the entire spike train if the bin width is chosen appropriately. After this first step, the dimension of the signal space is not uncountably large any more, but still infinite, so that further reductions will be necessary.

Remarking on the discretization of time, it is also possible to work with continuous spaces with countable dimension. Such spaces look somewhat more elegant because they allow to avoid the symmetry breaking with respect to time shifts. In case of time discretization, time shift invariance gets lost for time shifts that are smaller than the bin width of the time histogram. However, this kind of symmetry breaking does not matter too much, because per definition of the bin width, it has an effect on those temporal structures only, which are considered to be irrelevant.

3.3.2 Neuronal Memory, Reset, and the Representation of Time

The next obvious possibility to reduce the dimension of the signal space is to assume that the relevant memory of a single neuron is finite. This assumption follows directly from the fact that relevance in neural coding is defined on the basis of information processing tasks, which require consideration of a limited amount of history only. For illustration, humans are able to categorize flashed images very reliably [TFM96]. Clearly, the relevant memory of the neurons can not exceed the total response time of the subject, because for the decisioning in a classification task every dependence on the history preceeding the stimulus can only be noise. The observation that such tasks can be done quite rapidly implies that the neuronal response should depend on a rather small history of the synaptic inputs only.

In fact, the issue of neuronal memory turns out to be very important but has not received much attention until now. In a feed-forward network, the memory of the neurons of different layers adds up linearly. For illustration, if the network consists of 20 layers and the neuronal responses all depend on the last 50 milliseconds of their synaptic inputs, then the total response time is already as much as one second. Much more dramatic is the effect in recurrent networks, where the network memory need not decay at all. In conclusion, the fact that it is possible to set up experiments where the brain acts as an information processing device is not at all trivial, but requires some mechanism to clean the network memory.

The simplest possibility to reset the network state of a feed-forward network as fast as possible is to have neurons, for which the memory equals the temporal resolution. In that case, the total response time is equal to the sum of the neuronal response delays only. For illustration, an example of such a network is the one considered at the beginning of this chapter (Eq. 3.1). In such a network, the time course of the external variable is directly represented by the time course of the neuronal response (apart from a fixed delay). This idea of “faithful signal transmission” is the most interesting point underlying the idea of a *rate code*. Although rate coding has often been used as a strawman standing for low temporal resolution or the absence of interspike correlations, I believe that the proponents of rate codes rather think of rate codes in the sense that ’time represents time’ (cf. [Mov99]) with high temporal resolution (see section 11.2 for a detailed discussion).

Other interesting possibilities to reset the network require some global reset signal. In fact, shunting inhibition would be well-suited to this end [BPG99; Her]. In case the reset signal is applied periodically, one would speak of a neuronal clock or a clocked network. Another very intriguing possibility would be if the reset signal is triggered by the task. In that case, one could even think of different reset signals for different tasks that in a top-down manner adjust the network state to the particular task to be solved [BP01]. Clearly, these ideas are typically discussed in the context of selective attention.

3.3.3 Interspike Correlations

Since ’rate coding’ is often associated with Poisson neuron models, correlations between spikes are frequently used to establish ’temporal coding’ as an alternative to rate coding. As pointed out above, however, one should define a *rate code* by the interesting property that the neuronal memory equals the temporal resolution (’time represents time’), which clearly allows for several interspike correlations. In particular, coincident spikes in a population of neurons are rather a prediction of rate codes than an objection against them, if one takes a short absolute refractory period into account.

In any case, correlations are interesting only if they can be used to reduce the size of the signal space. The presence of interspike correlations *per se* are not very interesting. What is needed is an explicit model on how these correlations are generated and how they are used by subsequent neurons to read out the presynaptic signal.

In the second part of this thesis, it will be demonstrated that correlations between spikes may be disadvantageous even if they carry significant information. Therefore, one should be careful not to focus too much on the idea of maximizing the mutual information between stimulus and response, because it may be incompatible with other relevant demands on the neural code.

3.3.4 Redundancy and Computational Load

Further arguments used to constrain the set of candidate neuronal representations are the redundancy between neurons and the computational load. It is quite likely that there is a considerable amount of redundancy between different neurons, because neuronal information processing, in particular in the cortex, seems to be robust against the death of individual cells. Redundancy, however, can be implemented in different ways. A pure form of redundancy would be if each neuron has a number of copies that could do the same job if necessary. It is, however, also possible to think of redundant coding schemes, where each neuron still has a small independent contribution to the total representation, so that the accuracy will decrease with the loss of any neuron. In that case, there are still enough neurons, which are sufficiently similar so that the loss of a neuron would be tolerable. The effect of redundancy on the accuracy of a neuronal representation, will be part of the investigations presented in the first part of this thesis and the issue of robustness is addressed in chapter 10.

A similar criterion is the argument that the computational load of the individual neurons should be rather small (see chapter 7). In other words, a certain kind of neuronal representation is considered to be unlikely, if it requires computations from the individual neurons, which are too complex. An appropriate amount of computational complexity seems to be the *linear-nonlinear cascade model* [Chi01; SPPS04], where the multi-dimensional synaptic input is first mapped onto a scalar variable using a linear time-invariant filter, and subsequently the output of the neuron is determined through a nonlinear function of the scalar variable.

3.4 Population Coding

The neuronal representation of an external variable of interest typically requires a large population of neurons. Through theoretical studies of population coding, it is possible to explore how the response properties of individual neurons affect the representation of the entire population.

Historically, the notion of '*population coding*' has developed in a somewhat narrower sense, namely as the opposite to '*grandmother cell*', '*labeled-line*', or '*winner-take-all*' coding. These terms all denote the special case, when merely the label of the most active neuron is used to encode a certain variable of interest. The most prominent example of a population code is the color coding of the cones in the retina, of which is known that a combination of the graded activity of different sorts of cones is necessary for color discrimination.

Experimentally, population coding is frequently investigated by using the population vector method, which assumes that the value of the encoded variable is proportional to the weighted sum

$$\hat{\vec{f}} \propto \sum_{j=1}^{N} k_j \, \vec{f}_j \tag{3.53}$$

where the f_j denote the preferred values of the individual neurons $j = 1, \ldots, N$ and k_j denotes the current activation (i.e. the number of spikes in a certain time window) of neuron j. The relevance of this method has been impressively demonstrated in [LRS88] for neuronal representations of saccadic eye movements in the deep layers of the superior colliculus. Through reversible inactivation of small subsets of collicular neurons it has been shown that it is not the most active cell that exclusively controls the eye movement (as it has been hypothesized before), but rather the average preferred direction weighted by the activities of a large population of coarsely tuned neurons.

Nevertheless, the population vector method is only one of many possible population coding strategies, and actually, every encoding problem that cannot be decomposed into a set of sub-problems, where each sub-problem deals only with a single neuron, constitutes a *population coding problem.* In this sense, a grandmother cell coding scheme has to be considered a population coding strategy as well, because the winner-takes-all function induces strong correlations between the activation of different neurons. However, even more fundamental than statistical correlations between neuronal responses is the role of the loss function for a population code. Whenever the average risk is reduced by a number of neurons in such a way that the contributions of the individual neurons to the total risk reduction cannot be assessed separately, we are faced with a *population coding problem.*

In other words, the question whether we have to deal with population coding or with single cell coding, essentially depends on how we define the loss function or, loosely speaking, how we set up the coding problem. If one aims to figure out which stimulus aspects can be discriminated by a single cell, this constitutes by definition a single cell coding problem. Conversely, if one sets up the problem by starting with a certain external variable of interest and then seeks to determine those aspects of neuronal activity that contribute to the representation of this given variable, this will naturally lead to a population coding problem. These two complementary approaches may be illustrated with the complementary concepts of the *point spread function* and the *receptive field.* The point spread function aims to describe the patterns of neuronal activity caused by a dot stimulus (e.g. a bright pixel on a black screen), whereas the receptive field aims to describe the set of stimuli causing an increase of the firing rate of a given neuron [McI01].

3.5 Neuronal Encoding from First Principles

Many theoretical studies investigate certain features of neural codes by using an optimization approach. In particular, the efficient coding hypothesis has been expressed in terms of the *infomax principle*, which seeks to optimize the encoding in such a way that the mutual information (i.e. the log information gain, see below) between input (sensory signals) and output (neuronal response) takes a maximum. Due to the ecological motivation of this approach, the focus of research thereby is to approximate the correct prior statistics of sensory signals as closely as possible.

In the context of population coding there is another line of theoretical research, which makes use of the idea of optimal signal transmission. Models of optimal population coding, however, are usually not informed by empirical prior distributions of sensory signals, but rather seek to explore general encoding strategies that are most efficient to overcome the corruption of signals caused by certain types of neural noise.

In fact, efficient coding and population coding research are concerned with complementary aspects of signal transmission and have been investigated rather independently of each other. The following quote may be a good way to describe the relationship of both:

> *The efficient coding principle should not be confused with optimal compression (i.e. rate-distortion theory) or optimal estimation. In particular, it makes no mention of the accuracy with which the signals are represented [...]. This may be viewed either as an advantage (because one does not need to incorporate any assumption, or the cost of misrepresenting the input) or a limitation (because such costs are clearly relevant for real organisms).*
>
> Simoncelli & Olshausen [SO01]

It is true that information theory, which constitutes the basis of the efficient coding hypothesis, allows to study certain aspects of signal representations without the need to specify a particular loss function. Whether this is an advantage or a disadvantage depends, however, on how much the inclusion of the true loss function would change the results.

Here I am presenting population coding and efficient coding from a unifying point of view. In this way, it becomes clear what one might gain, if a model is additionally informed by knowledge about the loss function. Clearly, the efficient coding principle does not require to specify the cost of misrepresenting the input, only because this cost has already been specified by the fundamental assumptions of information theory, which is used by this principle. The next section presents the necessary basics of information theory in the context of optimal encoding.

3.5.1 Information Theory – Part II (encoding)

The fundamental problem that has lead to the invention of information theory by Shannon and Weaver [Sha48; SW49] may be sketched by the question for those conditions that are necessary and sufficient to send a given *source* over a given *channel* with **zero error**. For the sake of simplicity, we will only consider discrete time, memoryless sources and channels, such that the *source* is completely determined by a probability distribution $P(X)$ over the possible set of symbols $x \in X$, while the *channel* is given by a family of distributions $(P(K|f))_{f \in F}$, which for every possible input f specifies the probability to observe a particular output $k \in K$. Furthermore, X, F, K are supposed to be discrete sets[10] for the time being.

The task now is to seek for a *code* that maps each source symbol $x \in X$ to a **sequence** of input elements $(f_1(x), \dots, f_n(x))$. Whether a source can be send over a channel depends on whether one can find a *code*, which allows to transmit all symbols generated by the source sufficiently fast over the channel without error. Both, the probability distribution $P(X)$ of the source, as well as the shape of the channel, impose constraints on the possibility of error-free signal transmission. An important insight of classical information theory is the fact that these constraints can be investigated independently, so that one commonly distinguishes between the problem of *source coding* and the problem of *channel coding* [CT91].

Source coding is concerned with the fact that the plain index $|X|$, counting the total number of symbols in X, does not reflect the relevant constraint of the source on signal transmission because a code allows to map the source symbols to sequences of input elements of the channel. Instead, the *source coding theorem* tells us that the entropy of $P(X)$ reflects the essential complexity of the source, because it provides a tight lower and upper bound for the average length of the input sequence (i.e. the average *code word length*) that has to be transmitted over the channel.

In conclusion, discrete source coding deals with the minimization of the code word length necessary to describe the source completely. If the source distribution $P(X)$ is uniform, its entropy equals the log of the index $\log |X|$ of the support of $P(X)$, which is the maximum over all possible distributions with the same support such that no further compression is possible. Conversely, every source, which has not a uniform distribution exhibits some amount of (theoretical) *redundancy*

$$R[P(X)] \equiv \log |X| - H[X] \tag{3.54}$$

[10] Note that it is impossible to achieve error-free signal transmission in case X is a continuous set apart from the trivial case of noise-free signal transmission.

so that the problem of discrete source coding is frequently called *redundancy reduction*. While Eq. 3.54 defines the theoretical amount of redundancy as it is evaluated on the basis of the log loss, the operational definition is given by the average code word length of the code that is given by the source minus the minimal possible average code word length that can be achieved by using another code to describe the source signal.

Another reason, why the average code word length is often larger than actually necessary is that in practice, one often does not know the 'true' distribution $P(X)$. In that case, the average code word length is clearly given by $-E[\log \hat{P}(x)]$, which is the average log risk of the estimated distribution $\hat{P}(x)$. The additional bits that are required due to the error in the estimate of $P(X)$ are measured by the *Kullback-Leibler distance*

$$D_{KL}[P(X)||\hat{P}(X)] \equiv -E[\log \hat{P}(x)] - H[X] = E\left[\log \frac{P(x)}{\hat{P}(x)}\right] \quad . \tag{3.55}$$

In summary, coding theory provides an intuitive explanation for the meaning of the log loss as a measure of the average code word length. Redundancy reduction deals with the question of how to find a code with minimal average code word length, which can be bound from below by the entropy, which is the Bayes risk under log loss. Consequently, it is possible to give Bayesian uncertainty representations by 'subjective' (maximum entropy) distributions the following meaning: Loosely speaking, one could imagine that a Bayesian never looses hope that someone will come to tell him about the true value of the variable, about which he or she has only uncertain knowledge. To make it possible to get informed by someone else, it is still necessary, to announce a code for the different possible values. This leaves the Bayesian free to choose the code such that the most likely values should be encoded by preferably short code words. In other words, an operational definition for the distributions in Bayesian theory is given by the corresponding choice of a representation for the uncertain variable. This renders precisely the fact that the wronger ones prejudices are, the longer it will take to get informed about the truth.

The central measure in *channel coding* is the *mutual information*

$$I[f,K] \equiv \Delta[P(f,K), \log \hat{P}(x)] = H[f] - H[f|K] = E\left[\log \frac{P(f,K)}{P(f)P(K)}\right] \tag{3.56}$$

which is the information gain (see Eq. 3.32) under log loss. In contrast to the quadratic information gain, mutual information has the remarkable property to be symmetric in f and K, such that the following equalities hold:

$$I[P(f,K)] = H[K] - H[K|f] = H[f] + H[K] - H[f,K] \tag{3.57}$$

The amount of information that can be transmitted through the channel $(P(K|f))_{f\in F}$ depends on the distribution of the input $P(f)$. The maximum mutual information determined over all possible input distributions

$$C[P(K|F)] \equiv \sup_{P(f)} I[f,K] \tag{3.58}$$

reflects the *channel capacity*. This definition is justified because of the important *channel coding theorem*, which states that all rates below the capacity C are achievable. This means that a sequence of codes $f_n : X \to F^n$ exists, for which the maximal probability of error vanishes in the limit of large n. In conclusion, the condition $H[X] < C[P(K|F)]$ is necessary and sufficient for the possibility of error-free transmission of a source over a channel.

When generalizing information theory from the discrete to the continuous case, it is very important to distinguish the source coding problem from the channel coding problem. While there is a unique limiting value of mutual information for all sequences of partitions with decreasing maximum bin width, this is different for continuous sources. If X is continuous, the whole problem of redundancy reduction considered above becomes irrelevant, because the entropy (i.e. the minimal description length) of the elements of X diverges and hence error free signal transmission is impossible in limited time. While it is clear by the definition of the problem above that error-free signal transmission does not require to specify a cost function for possible errors, this becomes necessarily different in case of continuous distributions. In other words, a new issue arises in case of continuous sources, namely to decompose the degrees of freedom of the source variable into *source signal* and *source noise*. To this end, *rate-distortion theory* has been developed, where the issue of redundancy reduction is replaced with the more general concept of compression, which merely seeks to describe the **relevant** information of a source as efficiently as possible.

The elegance with which the formulas of discrete information theory can be extended to continuous sets is so seductive that this crucial difference between both concepts often remains obscured. Therefore, it should be emphasized that as soon as one departs from the goal to achieve error-free transmission, it is indispensable to make assumptions about the effective cost of an error [CT91]. In other words, there is no way around rate-distortion theory in case of continuous sources.

Nevertheless, models used in efficient coding and unsupervised learning research are typically constructed without a discussion of the loss function by resorting to the *infomax principle* [Lin88] instead. The infomax principle can be defined by the objective to maximize mutual information between a certain source and the output of a channel. At this point it is important to note that the infomax method is related to the channel coding problem only: **because of the separability of source coding from channel coding, it cannot tell anything about which part of information produced by the source should be disregarded!**

In conclusion, in such infomax models the decision, which information of the source is to be discarded is not explicitely controlled by the choice of a loss function, but is left to other ingredients of the model like, for example, the architecture of a neural network. Since these other ingredients are not suited directly to the source coding problem, it is difficult to assess which information exactly is discarded, and even more importantly, why this information is discarded.

One way to inform a model effectively about the distinction between signal and noise, without using a cost function, which has frequently been used in efficient coding, is to specify an appropriate generative model. Generally speaking, a *generative model* is a latent variable model of the source distribution

$$\hat{P}(x) = \sum_s \hat{P}(x|s)\hat{P}(s) \quad . \tag{3.59}$$

The latent variables s are intended to represent the relevant quantities (often called 'causes'), whereas the conditional distribution $\hat{P}(x|s)$ represents noise. While it is tempting to discuss generative models in terms of redundancy reduction [LO99] by minimizing the average log risk[11]

$$-E[\log \hat{\rho}(x)] = -E[\log \sum_s \hat{\rho}(x|s)\hat{P}(s)] \quad . \tag{3.60}$$

this is somewhat misleading, because it suggests that the major problem was to estimate the true distribution $P(X)$ (e.g. the 'natural scene statistics'), which would imply that all generative models are equivalent whenever they lead to the same optimal $\hat{P}(x)$. However, different latent variable models of the same distribution $P(X)$ make a big difference if one takes the maximum likelihood estimate $\hat{s}_{ML}(x)$ as the 'efficient representation' of the source signal x. Therefore, it is so important in efficient coding to use 'physically motivated' generative models, in which the latent variables rather reflect the behaviorally relevant parameters of a scene. Since the use of continuous variables for the description of these behaviorally relevant parameters corresponds to the intuition that something like the Euclidian metric reflects the relevant cost function, the knowledge incorporated by the choice of the generative model can be even better described by a loss function like $l(x,k) = (\hat{s}_{ML}(x) - \hat{x}_{MS}(x))^2$, which measures the error with respect to the relevant function of x.

In conclusion, generative models constitute an interesting alternative to inform a model how to distinguish signal from noise, but it is misleading to discuss continuous sources in terms of redundancy reduction, because its entropy (i.e. its minimal description length) diverges. Redundancy reduction makes sense only for sources with finite information rate. In particular, it is not true that *'one does not need to incorporate any assumption of misrepresenting the input'*, but the whole issue

[11] In case of continuous distributions the log loss is defined with respect to the density $\rho(x)$ of that distribution.

in unsupervised learning with continuous sources is about finding good reasons to prefer some information (i.e. the signal) to some other information (i.e. the noise). The most principled way (i.e. the most transparent way) to do so, is clearly to specify a loss function as it is done in rate-distortion theory.

3.5.2 Efficient Coding meets Population Coding

Summarizing the preceeding discussions on optimal decoding and encoding, there are three major sorts of knowledge, with which models of neuronal representations can be informed: *behavioral relevance*, the *relevant input statistics*, and *neuronal transmission noise*. Additionally, the limited *computational power of neurons* constitutes an important aspect too, which will be discussed in chapter 7.

The foremost problem is, in fact, to unravel the aspect of behavioral relevance because it determines the relevant input statistics, which is indispensable for coding. The investigations in this thesis, however, are not intended to contribute to this issue, but follow the approach in population coding, which is to resort to some abstract coding problem, in order to learn something about which encoding strategies are most effective to overcome the effect of neuronal noise.

While the focus of efficient coding research is to look rather for the relevant input statistics, the comparison of basis images with neuronal receptive fields relies on assumptions about the neural code as well. Therefore, it makes sense to relate efficient coding models to population coding. Concisely speaking, this thesis investigates efficient coding and population coding jointly with respect to the problem of channel coding (i.e. the problem of signal transmission), but it neglects the source coding issue, which is the actual core of efficient coding research.

A general scheme of signal transmission models is displayed in Fig. 3.7 showing the ingredients, which are necessary to evaluate the performance of an encoding strategy. The generative model is one way to specify the task. It is the task, which allows to consider a neuronal system as an information processing device, computing a function $\hat{s}(x)$ on the inputs x. The neuronal representation k of $\hat{s}(x)$ is likely to suffer from both, the input noise $P(x|s)$ as well as the noise of neuronal signal transmission $P(k|f)$. The quality of this representation also depends on the subsequent readout. An experimentalist tends to use the best possible readout, while in the brain, it should become more and more constraint at later stages of the system. Conversely, this means that at early systems of sensory processing, say for the striate cortex or at least for the retina, it makes sense to model the subsequent neuronal processing by a rather powerful estimator too.

Efficient coding models commonly use a loss function with a binary case distinction

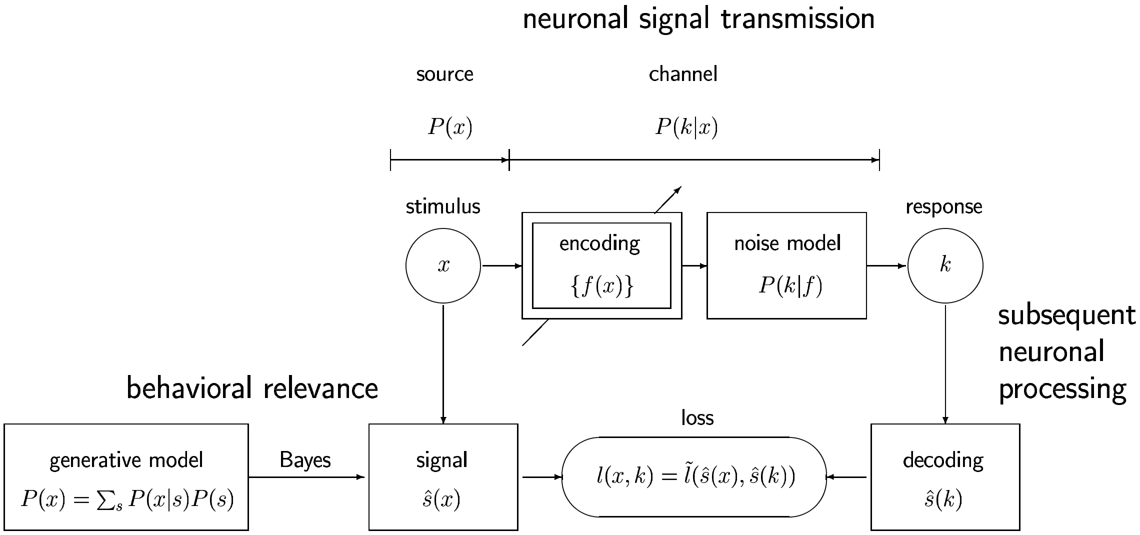

Figure 3.7. General scheme for deriving neuronal representations from first principles. The encoding is optimized by minimizing the average loss. The scheme brings to light the different sorts of assumptions that enter the optimization model: the source distribution $P(x)$, the loss function $l(x, y)$ which summarizes the distinction between signal $\hat{s}(x)$ and noise $P(x|s)$, the set of candidate encodings $\{f(x)\}$, the noise model $P(k|f)$, and the decoder $\hat{s}(k)$. Note that $\hat{s}(x)$ here may also constitute a representation of the posterior distribution $P(s|x)$.

$$l(x,k) = \begin{cases} \log \hat{P}(x|k) & , \quad if x \in \{natural\ stimuli\} \\ l_0 & , \quad otherwise \end{cases} \quad . \tag{3.61}$$

For natural stimuli it is the log loss, while otherwise it is constant. In this way, the task becomes equivalent to a stimulus reconstruction problem (i.e. $\hat{s}(x) = x$), when $P(x)$ is replaced with $P(x|x \in \{natural\ stimuli\})$. Efficient coding models typically resort to the infomax principle, which means that the decoder is not required to decide for a best guess $\hat{x}$, but it estimates a probability distribution $\hat{P}(X|k)$ instead. The decoder $\hat{P}(x|s)$ is supposed to be the optimal Bayes estimator, using $P(x|x \in \{natural\ stimuli\})$ as prior, so that it holds $\hat{P}(X|k) = P(X|k, x \in \{natural\ stimuli\})$. Simple models restrict the set of possible encodings to unitary mappings and assume constant additive noise, such that the optimal encoding makes the different dimensions of the output as independent as possible. These models constitute a linear version of the *independent component analysis (ICA).*

Population coding models stick with the stimulus reconstruction paradigm as well. In contrast to efficient coding, the distinction between relevant and irrelevant stimuli is not made explicit. Either certain stimulus parameters of interest supposed to be relevant are selected *ad hoc*, or the coding problem remains entirely abstract. Consequently, there is often no corresponding prior distribution that could be determined empirically.

Another difference is that most models in population coding demand from the decoder to make a decision for a best guess $\hat{x}$, which is commonly evaluated with respect to the squared error loss. The selection of the decoder has rather been a matter of taste: the most popular choices are the population vector method, maximum likelihood, Bayes, and linear Bayes estimators[12]. For the derivation of optimal population codes, however, all previous studies used the average Fisher information as objective function instead of evaluating the error of an estimator explicitly.

The standard noise model in population coding is Poisson noise, while the effect of using other noise models has been investigated quite extensively [Kar00; WE01]. In particular, several models of correlated noise have been studied [AD99; HHKM01].

A major goal in theoretical work on population coding is to determine signatures of advantageous encoding strategies. The encoding is commonly specified by an array of tuning functions, which determine the firing rate of each neuron as a function of the stimulus parameter(s) of interest. In particular, sets of candidate encodings that differ in the tuning widths have been investigated. This issue will be addressed in the next chapter.

Since a great deal of work in this thesis is about optimizing encoding strategies, I conclude this section with a warning, because optimization approaches carry a large risk to give wrong impressions about the assumptions used to obtain a certain result. While the first principle is typically emphasized showing up repeatedly in the literature, it actually constitutes only one of many aspects that are responsible for the results obtained from a model. First principles are discussed at length in the beginning and the end of a paper, but many relevant aspects of a model, other than the objective function, often remain hidden in the technical sections.

What is widely overlooked is the fact that also the problems typically studied in physics, say e.g. the electric field caused by charged metallic objects, is not at all determined by the first principle alone, but strongly depends on the particular constraints at hand, like e.g. the geometry of the objects. Therefore, a major goal of this thesis is to set up a common framework, which makes the assumptions of different studies in the literature explicitly comparable and to figure out whether and how results may change if one gambles a little bit around with the constraints that have been used so far.

[12] This estimator is presented in chapter 7

Part I

Optimal Neuronal Tuning

Chapter 4

Coarse Coding

– *Optimal population coding revisited*

One major motivation for the theoretical study of population codes originates from the problem, how to relate the shape of the tuning function of an individual neuron to the accuracy of a distributed representation, which makes use of the activity pattern of a large population of neurons. In particular, a lot of attention has been devoted to the problem how the tuning width affects the accuracy of population codes, since it was recurrently observed "*that the accuracy with which primates are able to perform perceptual or motor tasks is much better than expected from the tuning width of single cells that are presumed to be involved in these tasks*" [Vog90].

The expectation to have a simple correspondence between receptive field size and discriminability stems from the particular idea of a "grandmother cell" type of representation, where the label of the most active neuron encodes the current signal value (see Fig. 4.1). For population codes, the tuning width w clearly does not provide a lower bound for spatial resolution (see Fig. 4.2), and an early theoretical study by Hinton [Hin81; HMR86] demonstrates even a superiority of coarse coding: for *binary* radial symmetric tuning functions, distributed uniformly over a D dimensional stimulus space and *vanishing neuronal response variability*, the minimal decoding error scales according to

$$\sigma \propto w^{1-D} \quad . \tag{4.1}$$

This means that sharp and broad tuning are equally good in case of $D = 1$, while broad tuning is optimal for all $D \geq 2$.

Subsequently, several theoretical studies investigated the effect of the tuning width on the acuity of population codes. Eurich and Schwegler [ES97] confirmed the result given by Eq. 4.1 analyzing the same model as in [Hin81], while the infinitely sized stimulus space, was replaced with a D dimensional sphere. Another very different argument for the advantage of coarse coding was presented by Baldi and Heiligenberg

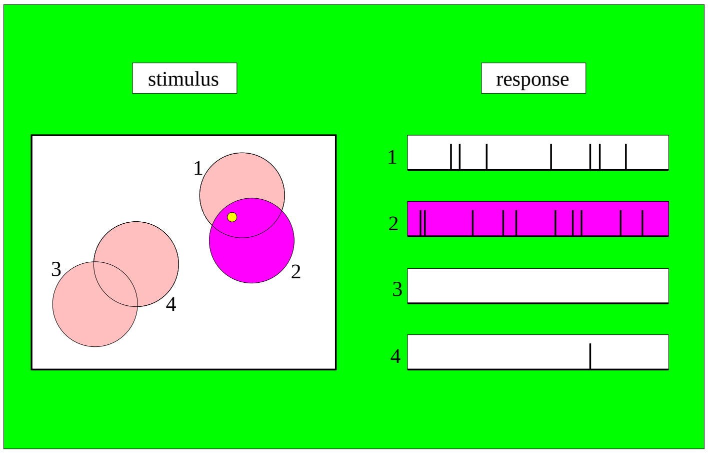

Figure 4.1. If the most active unit is the only one used to represent the stimulus, then the uncertainty about the location of the stimulus is directly related to the receptive field size of that neuron.

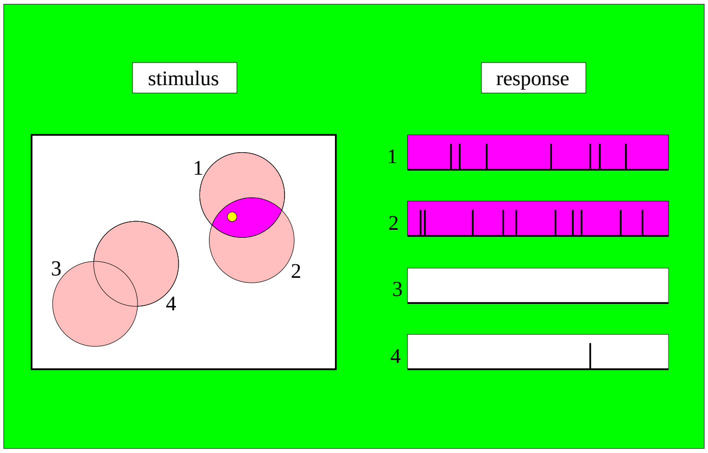

Figure 4.2. If more than one neuron is taken into account, the uncertainty about the location of the stimulus can be substantially smaller than the receptive field size of the individual neurons. If e.g. the combinatorics of all "sufficiently" active neurons is used, the uncertainty is rather related to the intersections of the receptive fields.

[BH88], who considered the approximation error of a particular radial basis function network.

On the other hand, Snippe and Koenderink [SK92] argued against the square nature of the sensitivity profiles used in [Hin81] and through their analysis using Gaussian receptive fields they found that coarse coding is not always optimal but that sharp tuning is better in the case of $D = 1$. While this conclusion contradicts the result of [BH88], which was based on Gaussian tuning curves as well, they claimed that their analysis was more accurate. More recently, the conclusion of [SK92] was supported by a work of Zhang and Sejnowski [ZS99], who derived an equivalent scaling rule on the basis of Fisher information

$$\sigma^2 \propto w_Z^{2-D}, \tag{4.2}$$

which they claimed to be "universal". In particular, they concluded for *any noise model* and *all* radial symmetric tuning functions, distributed uniformly over a D dimensional stimulus space, that sharp tuning is optimal in case of $D = 1$, sharp and broad tuning are equally good in case of $D = 2$, and broad tuning is optimal in case of $D \geq 3$.

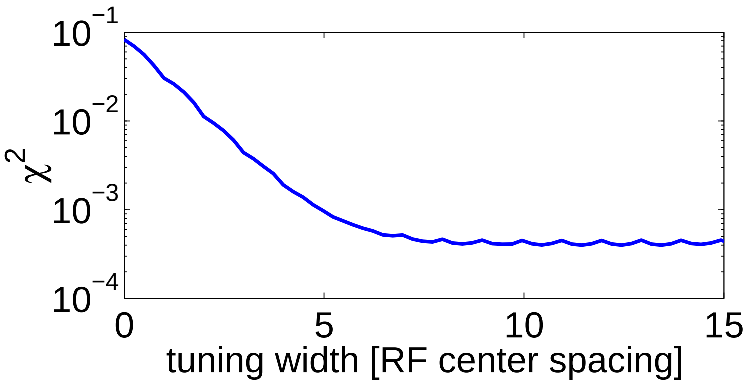

Figure 4.3. Minimum mean squared error as a function of the tuning width in case of $D = 1$ and a Poissonian spike count distribution with a maximum mean spike count equal to one. Furthermore, the model used has equidistantly spaced box tuning functions and a uniform prior distribution with a support of unit length.

Obviously, however, this “universal scaling rule” does not account for the particular case studied by Hinton. Since Fisher information is particularly relevant in case of small noise (see next chapter or [BRP02]), the absence of noise cannot explain this contradiction. In fact, Fig. 4.3 shows the numerical result that for $D = 1$, a small tuning width becomes even worse, if noise is increased. Furthermore, it will be shown that tuning functions exist, which are not binary, but exhibit the same scaling as predicted by Hinton's model. Hence, the latter does not constitute an irrelevant exception, but that shows a universal scaling rule cannot match Eq. 4.2.

Taken together, the literature on optimal tuning width did not provide a coherent picture and it was lacking a thorough discussion of the assumptions leading to divergent conclusions. In the following section, a new scaling rule is presented, according to which Eq. 4.1 and Eq. 4.2 can be understood as special cases, each of which representing a particular choice for the dynamic range of the tuning functions.

4.1 What determines the optimal tuning width?

The contradiction between both scaling rules, Eq. 4.1 and Eq. 4.2, can be resolved by recognizing that the ansatz for the variation of the tuning width used in [ZS99] implies a scaling of the dynamic range of the tuning function at the same time (see Fig. 4.4, left). In order to decouple the tuning width from the dynamic range of a tuning function, it is necessary to use a slightly different ansatz, which we will first demonstrate by a simple example, where the tuning function of each single neuron $f(\vec{x}) = f_{max}\phi(|\vec{x} - \vec{c}|)$ also depends only on the Euclidean distance to the center $\vec{c}$,

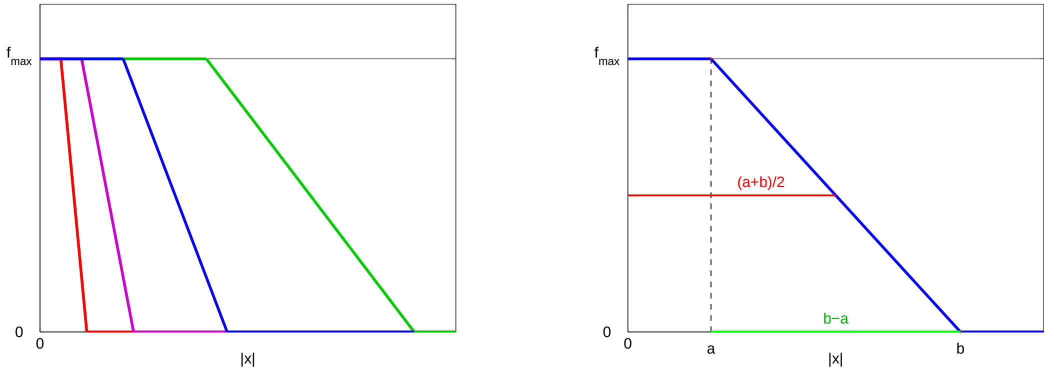

Figure 4.4. Radial component of the tuning functions. **Left:** Shows 4 examples for different values of the length scale parameter used in [ZS99] in order to illustration its simultaneous effect on both, the tuning width as well as the dynamic range. **Right:** Parameterization that allows to adjust the tuning width independent of the dynamic range.

but now can be adjusted by two independent parameters a and b

$$\phi(z) = \begin{cases} 1 & , \quad z < a \\ 1 - \frac{z-a}{b-a} & , \quad a < z < b \\ 0 & , \quad z > b \end{cases} \tag{4.3}$$

instead of a single scaling parameter only (see Fig. 4.4, right). Following [ZS99], we assume an (improper) uniform distribution of the tuning function centers $\vec{c}$ over the entire stimulus space by which the average tuning function array becomes isotropic and one may consider the reconstruction error w.r.t. an arbitrary direction, say $\vec{e}_1$, only. Furthermore, it suffices to consider the tuning function at $\vec{c} = 0$. In case of additive Gaussian noise with variance v the corresponding Fisher information component J_1 yields

$$J_1[f(\vec{x})] = f_{max}^2 \begin{cases} 0 & , \quad |\vec{x}| < a \\ \frac{1}{v} \cdot \frac{1}{(b-a)^2} \left(\frac{x_1}{|\vec{x}|} \right)^2 & , \quad a < |\vec{x}| < b \\ 0 & , \quad |\vec{x}| > b \end{cases} \quad . \tag{4.4}$$

The total Fisher information in the uniform neuron density approximation as used in [ZS99] is proportional to the average over J_1:

$$\bar{J} = \int J_1[f(\vec{x})] d\vec{x} \quad . \tag{4.5}$$

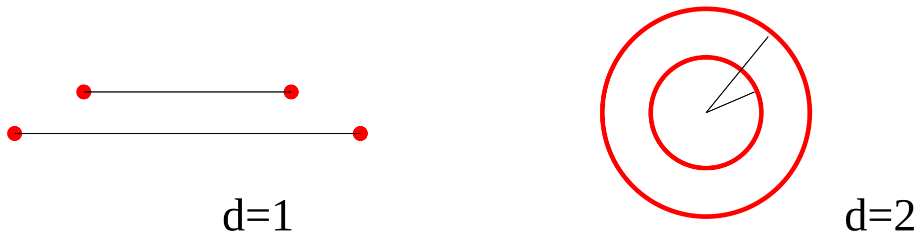

Figure 4.5. Geometry of coarse coding. For each tuning function the area of positive Fisher information corresponds to the surface of a d-dimensional sphere with radius w, which is proportional to w^{d-1}.

In case of $D = 1$ this yields

$$\bar{J} = 2 \cdot \frac{f_{max}^2}{v} \cdot \frac{1}{b-a} = 2 \cdot \frac{f_{max}^2}{v} \cdot \frac{1}{d} \quad , \tag{4.6}$$

where we defined $d := b - a$ as the dynamic range (i.e. the length of the region with positive Fisher information). In case of $D = 2$ we obtain

$$\bar{J} = \frac{\pi}{2} \cdot \frac{f_{max}^2}{v} \cdot \frac{b+a}{b-a} = \frac{\pi}{2} \cdot \frac{f_{max}^2}{v} \cdot \frac{w}{d} \quad , \tag{4.7}$$

where we defined $w := a + b$ as the tuning width. In case of $D = 3$ the average Fisher information reads

$$\bar{J} = \frac{4\pi}{9} \cdot \frac{f_{max}^2}{v} \cdot \frac{w^2}{d} \quad , \tag{4.8}$$

and for arbitrary D it is straightforward to prove the following scaling rule:

$$\bar{J} \propto \frac{f_{max}^2}{v} \cdot \frac{w^{D-1}}{d} \quad , \tag{4.9}$$

This scaling rule does not rely on the noise model as long as the amount of noise does not systematically depend on w or d, and it can simply be generalized to other tuning functions, where the ramp function, describing the radial component of the tuning curves, is replaced with some other decay profile (e.g. any sigmoidal function). As explained in [BRP02], however, the Fisher information $J_1(|\vec{x} - \vec{c}|)$ is tightly related to the minimum mean squared error only if the tuning functions are sufficiently regularized.

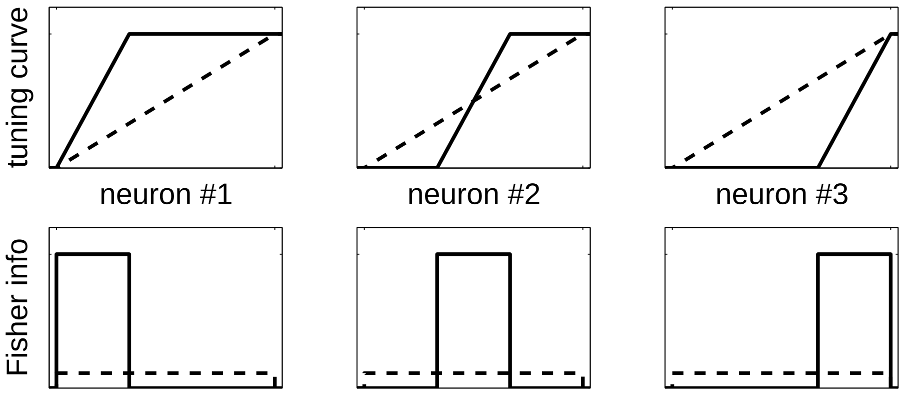

Figure 4.6. Why a small dynamic range increases total Fisher information. Coding scheme 1 (solid) with small dynamic ranges is compared with coding scheme 2 (dashed) with large dynamic ranges. The corresponding tuning curves are shown in the upper panel and the resulting Fisher information for each tuning function in case of additive Gaussian noise is shown in the lower panel. The total Fisher information of scheme 1 is three times larger than the total Fisher information of scheme 2

For a constant dynamic range Eq. 4.9 reproduces Hinton's scaling rule (Eq. 4.1). While the total Fisher information diverges in the limit $d \to 0$ due to the neuron density approximation, it will become clear in the next chapter how the effect of the dynamic range on Fisher information can be understood in detail. For the time being, one has to read this formula simply a rule of thumb for the scaling behavior, which also holds for small but not too small d.

Zhang's scaling rule (Eq. 4.2) is obtained under the special assumption $d \propto w$. In other words, Fisher information of radial symmetric tuning functions does not only depend on the width, but also on the dynamic range. The dependence on the width is well described by Hinton's rule and has the same simple geometrical explanation namely, that the surface of a D-dimensional sphere with diameter w is proportional to w^{D-1} (see Fig. 4.5). The additional dependence on the dynamic range is due to the fact that Fisher information is proportional to the squared derivative of the tuning function. Therefore, the contribution of a single tuning function to the total Fisher information decreases quadratically with an increase of the dynamic range so that the net effect for the total Fisher information is negative (see Fig. 4.6).

In conclusion, the long standing contradictions in the literature regarding the conclusions about the optimal tuning width have been resolved by making the underlying assumptions explicit. In this way, we discovered the steepness of a tuning function as a new, crucial parameter of population codes, which allows us to end up with the unique result that coarse coding is advantageous for all D.

It should be mentioned, however, that this result is clearly not the end of the story. There are several assumptions underlying the models considered so far that can be called into question. To name some important examples, we will see below that the advantage of coarse coding relies on . . .

- the radial symmetry of the tuning functions,
- the absence of energy constraints,
- and, in the context of vision, on the dot stimulus prior.

Furthermore, the use of Fisher information and the neuron density approximation to construct an objective function for the derivation of optimal population codes is rather awkward because as it turns out, those tuning functions that lead to a large total Fisher information are particularly likely to underestimate the true minimum mean squared error by far. This observation has been the motivation for a careful study of Fisher-optimal population codes, which will be presented in the next section.

Chapter 5

Fisher-Optimal Population Coding

The scaling rule for an optimal tuning width derived in the previous chapter is a good example for a model, which substantially relies on uncertain knowledge. Clearly, we have to ask, what the scaling rule can tell us about real neurons in the brain. Shall we now expect to find a large tuning width, whenever we measure a two-dimensional tuning function? Or can we falsify the efficient coding principle from the ubiquitous finding of smooth bell-shaped tuning functions, which are inefficient because of their large dynamic range?

While most people will not hesitate to consider these 'naive' conclusions invalid, it is, in turn not at all obvious which conclusions can actually be drawn. One of the foremost problems with the scaling rule is the question to which stimulus space it refers. Clearly, according to the efficient coding principle, neurons are not optimized to any arbitrary stimulus dimension (which would be an inconsistent demand), but one has to consider the naturally relevant stimulus space. This space, however, is unlikely to be selected by chance. Consequently, the requirement of radial symmetry or unimodality is very unlikely to hold with respect to the relevant stimulus space. Its ubiquitous finding in experimental data, however, rather reflects an experimental selection bias.

In conclusion, the assumption of a particular parameterization of a tuning function cannot be justified unless one knows about the space of all relevant stimulus parameters, which in particular cases might be possible. Therefore, we here depart from previous studies of optimal population coding, where the optimization of scale or width was based on a comparison between tuning functions only with exactly the same shape, neglecting the fact that the precision may depend much less on scale or width than on other aspects of the shape of the tuning functions. In contrast, we now seek to find population codes that are optimal within preferably large classes of tuning functions that are specified by very basic constraints only, like e.g. a limited maximum firing rate or unimodality. This means, in particular, that we do not restrict the class of encoding strategies *a priori* to only those tuning functions that are considered as biologically plausible, but we rather intend to check the actual explanatory power of Fisher-optimality as a first principle.

The relaxation of the restrictions on the set of candidate encodings, however, uncovers another, technical problem, which has been rather ignored in previous work, namely, that Fisher information tends to give meaningless results, the less regularized the space of encodings is. In fact, the hegemony of Fisher information as the commonly used measure of coding accuracy in the population coding literature is accompanied by a striking lack of justification. In particular, it seems that Fisher information has been more and more simply equated or even defined as *the* coding accuracy [PDL01], rather than treating it as an approximation or a minimalist calculus.

Since it turned out that the Fisher information approximation tends to fail, when used for the derivation of efficient encodings, the work presented in this chapter also emphasizes the technical question of when this approximation is able to give meaningful results. In spite of the restricted validity of Fisher information as a measure of coding accuracy, however, it is instructive to see how critically optimal encodings can depend on small changes in the assumptions. Furthermore, because Fisher information allows for an analytical treatment of population coding, it can serve as a tool to first develop hypotheses about optimal encoding strategies, which can be tested subsequently by numerical studies.

The chapter is organized as follows: In section 5.1, different ways to justify the use of Fisher information that are present in the literature, will be reviewed, and the most general argument based on asymptotic efficiency will be explained. In section 5.2 the optimal scale for the example of Gaussian tuning curves is determined with respect to Fisher information on the one hand and with respect to the minimum mean squared error (MMSE) in the case of a small counting time window on the other hand. This example indicates that Fisher information does not account for the MMSE in the case of short-term population coding. Subsequently, it is shown that this problem becomes especially relevant for Fisher-optimal codes if one drops the *a priori* restriction to Gaussian shaped tuning curves. This is demonstrated by presenting an example in section 5.3, where two neurons are sufficient to achieve arbitrary large Fisher information. The conditions under which Fisher information can be used to determine the MMSE are discussed in section 5.4. In section 5.5, we investigate the case, where the tuning functions are constrained to have a single maximum only (i.e. unimodal tuning functions) and the crucial role of energy constraints for the tuning width is demonstrated in section 5.6. Finally, the prerequisites and implications of the results are discussed in section 5.7.

Throughout this chapter the encoded random variable will be denoted by x and the observable spike count vector by $\mathbf{k}$, whose N components are the numbers of spikes of the N neurons (see Fig. 5.1). Because all observable quantities take values within

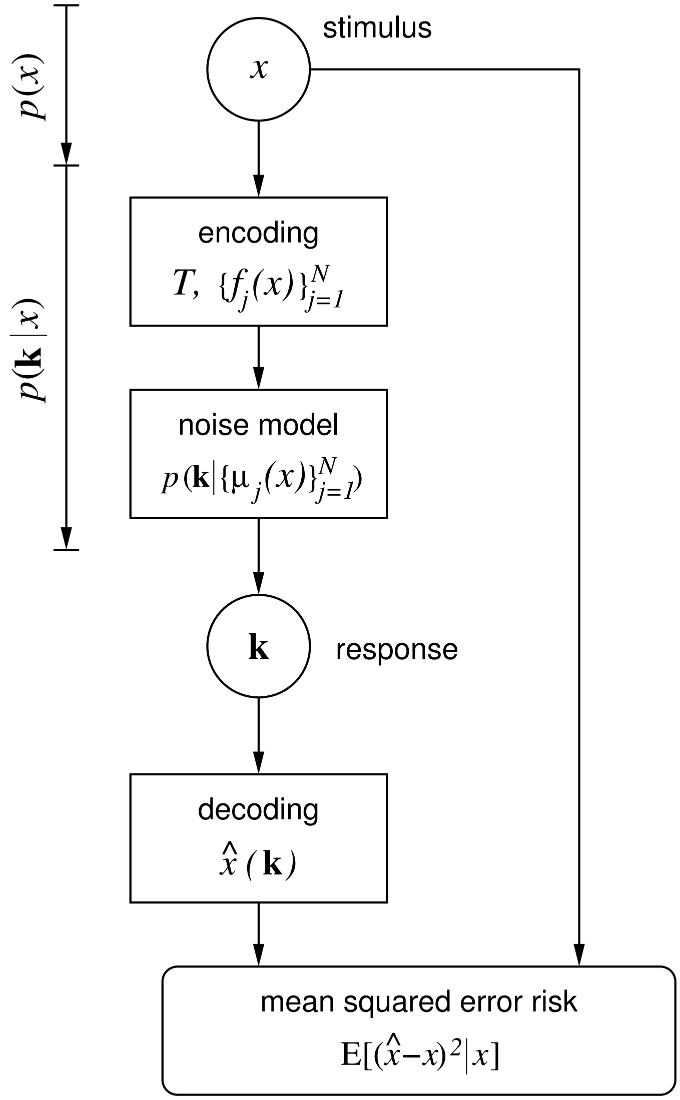

Figure 5.1. General population coding scheme. The relationship between a stimulus signal x and the neuronal response $\mathbf{k}$ is determined by the conditional probabilities $p(\mathbf{k}|x)$, which can be decomposed into the encoding and the noise model. The mean spike counts $\mu_j(x) \equiv \mathrm{E}[k_j|x] = Tf_j(x)$ as functions of x specify the encoding. Any function $\hat{x} : \mathbf{k} \mapsto \hat{x}(\mathbf{k})$ may be considered as a candidate estimator of x. The performance of an estimator $\hat{x}$ with respect to a given x is judged by its mean squared error risk.

a limited range only, we set the range of x to the open unit interval $x \in (0,1)$ without loss of generality. For convenience, we assume x to be uniformly distributed with density $p(x) = \Theta(x)\Theta(1-x)$, where $\Theta(y)$ denotes the Heaviside function, which is one, if $y > 0$, and zero otherwise. $p(x)$ is also called the *a priori* distribution, because it determines general properties of the signal x that are independent of the observed $\mathbf{k}$.

The encoding of x is specified by the set of tuning functions $\{f_j(x)\}_{j=1}^{N}$ that give the mean number of spikes for each neuron j divided by the length T of the counting time window. Together with the assumption of independent Poisson noise, the response statistics of the entire population is then described by

$$p(\mathbf{k}|x) = \prod_{j=1}^{N} p_{\mu_j}(k_j) = \prod_{j=1}^{N} \frac{(Tf_j(x))^{k_j}}{k_j!} \exp\{-Tf_j(x)\}\,, \tag{5.1}$$

where $p_{\mu_j}(k_j)$ denotes the probability mass function of the Poisson distribution with parameter $\mu_j = Tf_j(x)$, which gives the mean and the variance of the spike count. Apart from the asymptotic cases $T \to 0$ and $T \to \infty$, we will frequently consider the case $f_{max}T = 1$, i.e. each neuron does not fire more than one spike on average. As will be discussed in section 5.7, I suspect that $f_{max}T = 1$ is of the relevant order for signal transmission in cortex.

Throughout this chapter, we will consider the estimation of a single parameter only. In the case of multi-parameter estimation, the choice of the squared error distance is not sufficient to enable a well-posed comparison of different coding schemes, but additional specifications become necessary. While the optimization can be very complicated if the loss functional $\mathcal{L}[\hat{\mathbf{x}}, \mathbf{x}]$ depends on different dimensions in a non-linear fashion, it becomes rather simple if it is multi-linear, i.e.

$$\mathcal{L}[\hat{\mathbf{x}}, \mathbf{x}] = \mathcal{L}(\chi_1^2, \dots, \chi_D^2) = \sum_{d=1}^{D} c_d \chi_d^2 \, , \tag{5.2}$$

which makes the individual loss χ_d^2 of different parameters commensurable by an appropriate choice of weightings $\{c_d\}_{d=1}^{D}$. If one further assumes that no statistical dependencies between the different stimulus components exist (i.e. $p(\mathbf{x}) = \prod_{m=1}^{D} p(x_m)$), it is easy to relate the results below to the case, where many parameters, say D, have to be inferred simultaneously from the neuronal population activity[1]. In the absence of special constraints, the optimization problem reduces to the single parameter case, where optimal encoding is simply given by D subpopulations that encode each parameter independently and the number of neurons in each subpopulation has to be chosen, such that the contributions $c_d \chi_d$ to the total loss become equal. The more general case, where statistical dependencies between the variables exist, can often be traced back to the case without correlations, provided the weightings c_d are all identical. E.g. if $p(\mathrm{x})$ is given by an arbitrary multivariate normal distribution, for which all correlations are determined by the covariance matrix, one can always find another coordinate system $\tilde{\mathrm{x}}$ by the Karhunen-Loeve transformation [Jol86], for which all variables become independent $(p(\tilde{\mathbf{x}}) = \prod_{m=1}^{D} p(\tilde{x}_m))$.

In conclusion, the shape of optimal tuning functions is not necessarily related to the number of dimensions as the models of the previous chapter might suggest. Under specific assumptions, however, dependencies on the number of dimensions can emerge. In particular, the scaling rules presented in the previous chapter, are a direct consequence of the restriction to radial symmetric tuning functions.

[1] This corresponds to the case considered in the previous chapter 4 apart from the assumption of radial symmetry.

5.1 How to justify the use of Fisher information in population coding

The first publication in population coding using Fisher information I am aware of was the work of Paradiso in 1988 [Par88], which relates the psychophysical performance of orientation discrimination to the Cramer-Rao bound, where the Fisher information has been derived from a statistical model of a cortical hypercolumn. In that paper the Cramer-Rao bound is printed only for the special case of unbiased estimators

$$\mathrm{Var}\left[\hat{x}\right] \geq \frac{1}{J[p(\mathbf{k}|x)]} \quad , \tag{5.3}$$

whereas in the text it is stated inconsistently that the inequality holds for '*any* estimate'. Apart from the Cramer-Rao bound, Paradiso also alluded to the asymptotic behavior of the ML estimator. The original statement 'If a large number of cells are used for the orientation estimation, this limit is attainable (in the sense that the variance of a maximum likelihood estimate of orientation asymptotically equals the lower bound)' again is too sloppy as it does not mention the necessary conditions, under which this statement holds true. Special care is required, however, because tuning functions of different neurons are typically not identical, so that the limit of a large number of neurons is not equivalent to the standard case of the limit of a large number of i.i.d. samples. Because of this departure from the standard case, it is indispensable to justify the use of Fisher information in a more explicit, thorough way.

The use of Fisher information as an asymptotic approximation of the true mean squared error is typically innocuous for the special case of smooth unimodal tuning functions, **provided they are independent of the total number of neurons** (This case has been studied e.g. by Paradiso [Par88] or Seung and Sompolinsky [SS93]). Optimal encodings, however, typically depend on the total number of neurons. For example, if we consider the model studied by Zhang and Sejnowski [ZS99], the optimal scale of the tuning functions should be as small as possible in case of $D = 1$. As small as possible means that the scale decreases proportionally to $1/N$ with increasing number of neurons N. It will be shown below that in this case no asymptotically efficient estimator exists with respect to the limit $N \to \infty$! This demonstrates that the optimization of population codes with respect to Fisher information is particularly awkward. Previous papers, however, where Fisher information has been used to look for optimal encodings [ZS99; EW00], do not mention this problem, but justify the use of Fisher information by alluding merely to the Cramer-Rao bound again.

Apart from the problem that the Cramer-Rao bound is known to be not sharp [LC99], another issue arises in the context of Fisher-optimal encodings, which is in need of clarification. A restriction to unbiased or asymptotically unbiased estimators can be a reasonable strategy in case of looking for optimal *de*coding, when a UMVU estimator exists. However, it is easy to see that there is no uniform maximum Fisher information (UMFI) encoding because without regularization of the encodings, Fisher information can be made infinitely large at any singular point. In other words, even if one accepts a restriction to unbiased or asymptotically unbiased estimators and supposing that the risk of one of these estimators can be approximated by the inverse Fisher information, this is still not sufficient to determine a unique optimal encoding with respect to the risk function.

In [ZS99; EW00] this problem is obscured by the use of the uniform neuron density approximation, which artificially ensures that the approximate population Fisher information does not depend on the stimulus. However, the neuron density approximation is justified only for those encodings, for which the *true* total Fisher information of the population is constant. If the true total Fisher information is constant, however, there is no need anymore to introduce an approximation. In conclusion, nothing is achieved by the use of the neuron density approximation, except that it misleads the reader to overlook the limited validity of the approach. Additionally, the wrong impression about the alleged generality of this approach is supported in [ZS99] by the repeated emphasis on the generality of the 'universal scaling rule', which 'includes all radial symmetric tuning functions'.

It is very seductive to accept the introduction of a density approximation, because there are so many successful examples in physics, where this leads to valid predictions. This is, however, not the case in optimal population coding. The meaning of the objective function, in particular, should not rely on this specific approximation.

5.1.1 The MMSE as Objective Function

A transparent way to construct an objective function in a meaningful way is to specify a total ordering on the risk functions as it is common in estimation theory and only then to ask for the conditions under which the inverse Fisher information can be used to approximate the risk function of certain estimators.

As mentioned above, the restriction to unbiased estimators does not lead to a total ordering on the risk functions, if the *en*coding is optimized instead of the *de*coding[2]. In fact, however, there is hardly any good reason to not accept the Bayesian approach, in the case of optimizing the *en*coding for optimal signal transmission. Although some researchers in population coding tend to avoid Bayesian methods, one should recognize at least its unquestioned hegemony in information engineering.

In signal processing, the minimax approach is particularly appropriate, because it

[2] While also in the case of decoding the assumption of unbiasedness does not always allow for determining a uniformly best estimator, this problem is completely devastating in case of encoding.

ensures minimal accuracy for all signals and not only on average. Correspondingly, this criterion would lead to a maximin approach with respect to the Fisher information approximation.

Similar to what is frequently done in minimax estimation, I will begin, however, with the average risk criterion and then I derive constant Fisher information encodings (analog to constant risk decodings) by using a uniform prior (which is analog to the least favorable distribution). There are some more reasons, why the average risk criterion is preferable in this context. First, because the Bayes risk can be seen as an ultimate limit for the coding accuracy, which, in particular, allows to reveal the limits of the Fisher information approximation: if for any prior the minimum mean squared error (MMSE) is larger than the inverse Fisher information, this is clearly sufficient to prove that no estimator can exist with a risk function equal to the inverse Fisher information. Furthermore, the average risk criterion

$$\chi^2 = \mathrm{E}[(\hat{x}-x)^2] = \mathrm{E}[\mathrm{E}[(\hat{x}-x)^2|x]] = \mathrm{E}[\mathrm{E}[(\hat{x}-x)^2|\mathbf{k}]] \tag{5.4}$$

is computationally much more convenient than the minimax error, and last but not least, it constitutes the standard distortion measure in rate-distortion theory [CT91].

Recall that the best estimator $\hat{x}_{MS}$, which minimizes χ^2, is given by the MS-estimator

$$\hat{x}_{MS}(\mathbf{k}) = \mathrm{E}[x|\mathbf{k}] = \int x\, p(x|\mathbf{k}) dx\,. \tag{5.5}$$

and its average risk, the MMSE in general reads

$$\chi^2_{MS} = \mathrm{E}[(\hat{x}_{MS}-x)^2] = \mathrm{E}[x^2] - \mathrm{E}[\hat{x}^2_{MS}]\,. \tag{5.6}$$

In the next section, it will be explained under which conditions the risk of this estimator can be approximated by the inverse Fisher information.

5.1.2 Asymptotic efficiency

The essential justification for the use of Fisher information is the concept of asymptotic efficiency. If one considers sequences of estimators $(\hat{x}_m)_{m=1}^{\infty}$, for which each element $\hat{x}_m(\mathbf{k}(1), \ldots, \mathbf{k}(m))$ refers to m independent, identically distributed (i.i.d.) spike count vectors $(\mathbf{k}(1), \ldots, \mathbf{k}(m))$, the corresponding sequence of risk functions asymptotically decreases proportionally to $1/m$ for many types of estimators[3]. This can essentially be explained by the central limit theorem. More precisely, it can be shown under some rather weak assumptions about $p(\mathbf{k}|x)$ and $p(x)$ (for details see

[3] 'Type of estimator', which is often called just 'estimator' as well, denotes a unique construction rule for estimators like the 'maximum likelihood' or 'minimum mean squared error' method.

[LC99]) that the rescaled error $\sqrt{m}(\hat{x}_m - x)$ converges in law to a normal distribution with zero mean and a variance, which is the reciprocal value of Fisher information

$$J[p(\mathbf{k}|x)] \equiv \mathrm{E}[(\partial_x \log p(\mathbf{k}|x))^2|x]\,. \tag{5.7}$$

Such (sequences of) estimators are called *asymptotically efficient*.

In the case of a homogeneous Poisson process, as considered here, it is equivalent to consider sequences of increasing decoding time windows $T_m = mT_0$. While the corresponding estimators are functions of a single spike count vector $\mathbf{k}$ only, this spike count vector can be interpreted as the sum $\sum_{t=1}^m \mathbf{k}(t)$ of m spike count vectors that are independently drawn from a Poisson distribution, corresponding to the time window T_0. Thereby no information gets lost, because $\sum_{t=1}^m k_j(t)$ is a sufficient statistics for the parameter of the Poisson distribution [LC99]. Accordingly, for asymptotically efficient estimators holds

$$\lim_{T\to\infty} r(x,T) \cdot J[\{f_j(x)\}_{j=1}^N] = 1\,, \tag{5.8}$$

where $r(x,T)$ denotes the risk depending on T and the Fisher information is determined by

$$J[\{f_j(x)\}_{j=1}^N] = T \sum_{j=1}^N \frac{f_j'^2(x)}{f_j(x)}\,, \tag{5.9}$$

which is obtained by inserting Eq. 5.1 into Eq. 5.7 [Par88; SS93].

While Fisher information also shows up in a similar way in the Cramer-Rao bound for unbiased estimators, the latter by itself is not sufficient to justify the use of Fisher information as a general measure for coding precision. In its general version, the Cramer-Rao bound leads to a lower bound for the risk of any estimator

$$r(x) \geq \frac{(\partial_x \mathrm{E}[\hat{x}(\mathbf{k})|x])^2}{J[p(\mathbf{k}|x)]} + (\mathrm{E}[\hat{x}(\mathbf{k})|x] - x)^2\,, \tag{5.10}$$

which is not unique for different estimators. Furthermore, even if uniqueness is given as e.g. in case of uniformly unbiased estimators, a comparison of different encodings can not be traced back to a comparison of some lower bounds on the decoding risks. Instead, it is indispensable to determine a sufficiently close *approximation* of the actual values. Since the exact equality $r(x) = J(x)^{-1}$ holds true only in very rare cases (see appendix 5.8.1), the notion of asymptotic efficiency presented above is crucial for the use of Fisher information[4].

5.1.3 MMSE and Fisher information

Provided a set of regularity conditions hold, the MS-estimator is known to be asymptotically efficient [LC99]. As explained above, this means that with an increasing

[4] For a more detailed discussion of differences and relationships between asymptotic theory and the Cramer-Rao bound see [LC99]

number of observations its risk asymptotically approaches $1/J(x)$ for all x. Because this of course implies that also the mean values of both converge, the MMSE is then asymptotically equal to the *mean asymptotic error (MASE)*

$$\chi^2_{AS} \equiv E\left[\frac{1}{J[\{f_j(x)\}_{j=1}^N]}\right] = \frac{1}{T}\int_0^1 \left(\sum_{j=1}^N \frac{f_j'^2(x)}{f_j(x)}\right)^{-1} dx\,. \tag{5.11}$$

Note, however, that the finite number of neurons is crucial for the accuracy of population codes, so that this limiting behavior is only useful as a guide of intuition, but not in a rigorous sense. This means that, even if one can prove asymptotic normality for a certain sequence of estimators $(\hat{x}_m)_{m=1}^\infty$, one still has to figure out, *how large* at least m has to be so that the MASE becomes a good approximation of the MMSE (i.e. for the relative difference between both it holds $|\chi^2_{MS} - \chi^2_{AS}|/\chi^2_{MS} < \epsilon \ll 1$). For example, it will be demonstrated below that T typically has a large effect on the shape of the MMSE-optimal code, although as explained above the MS estimator is asymptotically efficient with respect to $T \to \infty$. In contrast, the shape of Fisher-optimal codes are necessarily independent of the available decoding time because T appears only as a constant factor in Eq. 5.11.

Previous papers on population coding, using Fisher information, referred rather to the limit of large N than to the limit of large T considered above. There is an important difference between these two kind of limiting processes: As long as the tuning functions are taken to be static, the integration of spikes over time corresponds to a sum over i.i.d. spike count vectors. In contrast, for the spike counts of different neurons we typically have $p(k_i|x) \neq p(k_j|x)$ for all $i \neq j$. This diversity of the tuning functions can crucially slow down the convergence of the MMSE to the MASE or it may even destroy the property of asymptotic efficiency. In fact, it is possible to construct sequences of tuning functions so that the difference between the MASE and MMSE becomes larger and larger the more tuning functions are taken into account by the MS-estimator (an example will be given below). Furthermore, as it turns out, those tuning functions that lead to large Fisher information are particularly likely to underestimate the MMSE by far. Clearly, the latter becomes a severe problem, when Fisher information is used as an objective function in order to determine optimal encodings.

Another fundamental problem is that Fisher information can be used for those encodings only, for which x is *identifiable*. This means that the mapping of the tuning functions has to be one-to-one. If x is not identifiable, Fisher information may either underestimate or overestimate the true error by far. E.g., a single symmetric tuning function centered at $1/2$ (the middle of the interval $(0, 1)$) cannot improve the mean squared error at all for any x, while Fisher information can be arbitrarily large everywhere. Conversely, the Fisher information of a single tuning curve that is constant somewhere within an arbitrary small, but finite interval, predicts a diverging error within this interval, while the risk of the MS-estimator, in fact, depends on the length of this interval and χ^2_{MS} can never be larger than the variance of the *a priori* distribution (Eq.5.6). In conclusion then, the use of Fisher information can at most be justified for the restricted set of encodings, for which the mapping of the tuning function array is one-to-one and hence, *Fisher-optimality* is here defined as to require both, a minimal MASE as well as identifiability.

5.1.4 Alternative Justifications

The justification of Eq. 5.11 as an objective function constructed from Fisher information as outlined above is actually the most general that is present in the literature. Nevertheless, one of our referees gave us a hard time with publishing [BRP02] because he rather believed that the limited validity of Fisher information as a measure for population codes uncovered by this justification, is a deficiency of the choice of the MMSE as objective function and not due to Fisher information itself. This is clearly not true. In fact, it should be noted that considering only the mean of the risk function (instead of its entire shape) is a rather graceful way, which at most underestimates the limitations of the Fisher information approximation! Once more, if for *any* prior the minimum mean squared error (MMSE) is larger than the inverse Fisher information, this is indeed sufficient to prove that no estimator can exist with a risk function equal to the inverse Fisher information.

Conversely, in the cases where one drops the common restriction to smooth, bell-shaped tuning functions like Gaussian or cosine tuning profiles, it is also possible that the MMSE becomes smaller than the mean inverse Fisher information[5]. In these cases, unbiased estimators – if they exist at all – are strongly suboptimal not only with respect to the average risk but also with respect to the maximum risk. While it is a wide-spread believe that unbiasedness is a generally desirable property this is not true even from a purely frequentist's point of view.

In conclusion, if one refers to the squared error loss, the (approximate) equality of the MMSE with the MASE is a **necessary** condition for a meaningful use of Fisher information.

Another line of motivation for the use of Fisher information originates from the finding that Fisher information also shows up as the leading coefficient in a Taylor expansion of the Kullback-Leibler distance:

$$D_{KL}(P(k|x)||P(k|x+\Delta x)) = J[P(k|x)]\frac{\Delta x^2}{2} + \mathcal{O}(x^3) \tag{5.12}$$

which builds the basis for *information geometry* [AN00] and has interesting applications in the context of neural coding [WNiA01]. In fact, I think the interpretation of

[5] This is the case if Fisher information is small within a surrounding of a stimulus parameter value, which preferably happens, if the tuning functions are flat within large regions like for example for sufficiently box-shaped tuning functions.

Fisher information as a ’rate’ of the Kullback-Leibler distance is the most intuitive one. As such, however, it is obvious that Fisher information is neither sufficient to serve as a reliable measure for the mean squared error coding cost nor to determine the mutual information of an encoding.

5.1.5 Fisher information, MMSE, and Mutual Information

In the population coding literature, the accuracy of a code is frequently simply equated with its Fisher information. In fact, it appears that it is sometimes taken as a definition rather than as a means to approximate the coding error [PDL01]. It is important to note, however, that ’Fisher information’ does not always make sense as an information measure although it carries the word ’information’ in its name.

The paper [BN98], which relates Fisher information to mutual information, already contains several hints that Fisher information gives unreasonable results in case of finite size/finite time conditions. I am afraid, however, that it has rather strengthened the view to consider Fisher information as an information measure of population codes *per se* and not as an approximation.

The relationship between Fisher information and mutual information established in [BN98] builds upon the notion of asymptotic normality as well. It is important to note, however, that the whole point in the asymptotic relationship between these two information measures is nothing but the distinct role of the normal distribution due to the central limit theorem. In order to illustrate this fact, consider some arbitrary feature of the distribution, say for example the *Fritze information*[6], which is defined to be the supremum of a density in case of continuous distributions:

$$H_F[\rho(x)] \equiv \sup_x \rho(x) \quad . \tag{5.13}$$

Since the variance as well as the entropy of a Gaussian is uniquely determined by its maximum, it is clearly possible to relate the Fritze information to mutual information in the case of asymptotic normality. In that case it holds:

$$I[X,Y] = H[X] - \mathrm{E}\left[\log\left(\frac{2\pi\sqrt{e}}{H_F[\rho(x|y)]}\right)\right] . \tag{5.14}$$

Furthermore, the Fritze information imposes an interesting lower bound on the average risk under 0-1-loss for *any* estimate:

$$\lim_{\epsilon \to 0} \frac{\mathrm{E}\left[l_\epsilon(x,\hat{x})\right]}{\epsilon} \geq \mathrm{E}\left[H_F[\rho(x|k)]\right] \quad . \tag{5.15}$$

[6] The origin of the name is due to the German tongue twister:“Fischers Fritze fischt frische Fische...”

and the bound is always attained by the MAP estimator. I hope this small humorous example helps to better assess the meaning of asymptotic relationships and general lower bounds.

Now, we turn to the question, how the three quantities, Fisher information, mutual information, and the MMSE are related to each other also beyond the asymptotic limit. In order to address this question let us consider the following bound

$$\begin{aligned}
\mathrm{I}[\mathbf{X},\mathbf{K}] &= H[X] - \mathrm{E}\left[H[X|k]\right] \\
&\geq H[X] - \mathrm{E}\left[\frac{1}{2}\log\left(2\pi\, e\ \mathrm{Var}\left[X|\, k\right]\right)\right] \qquad (5.16)\\
&\geq H[X] - \frac{1}{2}\log\left(2\pi\, e\ \mathrm{E}\left[\mathrm{Var}\left[X|\, k\right]\right]\right) \qquad (5.17)\\
&= H[X] - \frac{1}{2}\log\left(2\pi\, e\ \chi^2_{MS}\right) \qquad (5.18)
\end{aligned}$$

for which in (5.16) equality holds in case of the Gaussian channel and hence, in the asymptotic normal case as well. The second bound (5.17) is due to Jensen's inequality and hence, equality holds if and only if $\mathrm{Var}\left[X|\, k\right]$ is constant almost everywhere. Clearly, the resulting inequality (5.18) can also be rewritten as a lower bound for the MMSE:

$$\chi^2_{MS} \geq \frac{\exp\{2(I[X,K] - H[X])\}}{2\pi\, e} \qquad (5.19)$$

Similar to the definition

$$I_{Fisher} = H[X] - \mathrm{E}\left[\frac{1}{2}\log\left(\frac{2\pi\, e}{J(X)}\right)\right] \qquad (5.20)$$

used in [BN98] one may define

$$I_{MS} = H[X] - \frac{1}{2}\log\left(2\pi\, e\ \chi^2_{MS}\right) \qquad (5.21)$$

and

$$I_{AS} = H[X] - \frac{1}{2}\log\left(2\pi\, e\ \chi^2_{AS}\right) . \qquad (5.22)$$

The only difference between I_{AS} and I_{Fisher} is the Jensen inequality. Therefore, it holds $I_{AS} \leq I_{Fisher}$ and equality holds if and only if Fisher information is constant almost everywhere. Since the Fisher-optimal codes, which are going to be derived in the following have indeed constant Fisher information, the use of χ^2_{AS} as objective function is equivalent to the use of I_{Fisher}.

In case of a single tuning curve and the Poisson noise model, Fisher information is constant for a quadratic tuning function $f(x) = f_{\max}x^2$ (see Fig. 5.4, left). The

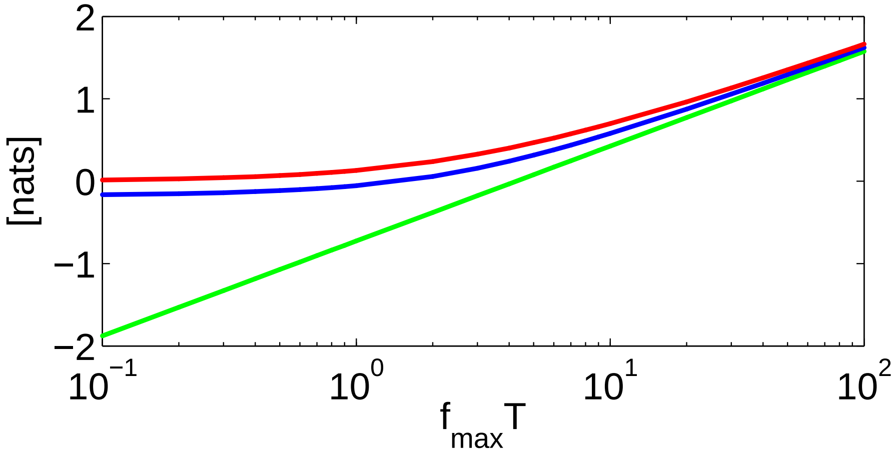

Figure 5.2. Comparison of I_{MS} (blue), I_{AS} (green), and the mutual information $I[X, K]$ (red) as a function of $f_{\max}T$.

resulting courses of I_{MS}, I_{AS}, and the mutual information $I[X, K]$ are shown in Fig. 5.2 as a function of T. Here, the MMSE and mutual information behave very similarly, while Fisher information is rather different. This observation demonstrates that Fisher information is only approximately valid also with respect to mutual information.

A very instructive example is given by the following tuning function

$$f_\omega(x) = f_{\max}\left[(\omega x) \bmod 1\right]^2 . \tag{5.23}$$

In case of $\omega = 1$, it is just the quadratic tuning function considered above. With increasing ω, however, the tuning function constitutes a wave function with increasing wave number ω (see Fig. 5.4, right). The interesting thing about this type of tuning function is that all three measures, mutual information, MMSE, and Fisher information take different values in the limit $\omega \to \infty$. While the MMSE increases with ω, converging to the *a priori* error $\mathrm{Var}[X]$ (i.e. no quadratic information gain at all), Fisher information diverges as it holds $J(x) = 4f_{\max}T\omega^2$ in this case. Mutual information, however, is identical for all ω.

This example also demonstrates nicely the difference between channel coding and rate distortion theory. To make this point as clear as possible think of $x = \sum_{j=1}^{\infty} b_j 2^{-j}$ to be represented by the binary sequence $(b_j)_{j=1}^{\infty}$. Furthermore let us suppose a noise model such that only the finite string of the first m elements (i.e. $(b_j)_{j=1}^{m}$) is observable. If we now compare, say $f_1(x) = x$ with $f_2(x) = 2x \bmod 1$, then in both cases m bits are transmitted and hence the mutual information is $\log_2 m$ bits. However, the kind of information is not the same: while the string $(b_j)_{j=1}^{m}$ is transmitted in case of f_1, it is the string $(b_j)_{j=2}^{m+1}$ transmitted in case of f_2. More generally, if we define $f_n(x) = 2^n \bmod 1$, the transmitted string $(b_j)_{j=1+n}^{m+n}$ is completely independent from $(b_j)_{j=1}^m$ if $n \geq m$.

The purpose of a loss function is to introduce a weighting or ranking which kind of information is most worthwhile for the problem at hand. In our example the quadratic loss corresponds to a weighting, which decays exponentially as 2^{-j}. This means the first bit is twice as much important as the second bit and so on. Since we typically have this idea in mind, when dealing with continuous variables, the MMSE is often more appropriate as objective function in case of encoding continuous variables than mutual information.

In conclusion, the channel determines only *how much* information can be transmitted. If the entropy rate of the source is larger than the mutual information of the channel, the issue of encoding is mainly an issue of choice, *which* information is to be selected to be transmitted.

5.2 Optimal Gaussian tuning depends on available decoding time

In this section, we start with a reinvestigation of Fisher-optimal encoding strategies in case of one-dimensional Gaussian tuning tuning functions under the assumption of a Poisson noise model. The MASE is compared with the MMSE for a given, limited number of neurons N ($N = 10$ and $N = 100$) and a finite decoding time window of length T. Similar to previous studies by Panzeri et al. [PBR$^+$96; PTSR99], we will focus on short time windows, since psychophysical experiments have shown that efficient computations can be performed in cortex at a rate where each neuron has fired on average only once [RT94; TFM96]. Precisely speaking, we consider the example of equidistant Gaussian tuning curves on the unit interval

$$f_j^{Gauss}(x) = f_{max} \exp\left\{ -\frac{1}{2}\left(\frac{x - j/N}{\sigma}\right)^2 \right\} \quad , j = 1, \ldots, N \, . \tag{5.24}$$

In order to determine the optimal scale with respect to the MASE, the corresponding Fisher information is calculated by inserting Eq.5.24 into Eq.5.9

$$J[\{f_j^{Gauss}(x)\}_{j=1}^N] = \frac{T f_{max}}{\sigma^2} \sum_{j=1}^{N} \left(\frac{x - j/N}{\sigma}\right)^2 \exp\left\{ -\frac{1}{2}\left(\frac{x - j/N}{\sigma}\right)^2 \right\} . \tag{5.25}$$

Numerical integration over $1/J[\{f_j^{Gauss}(x)\}_{j=1}^N]$ then yields the MASE $(\chi^2_{AS})^{Gauss}$. If it is multiplied by $f_{max}T$, the resulting expression becomes independent of time,

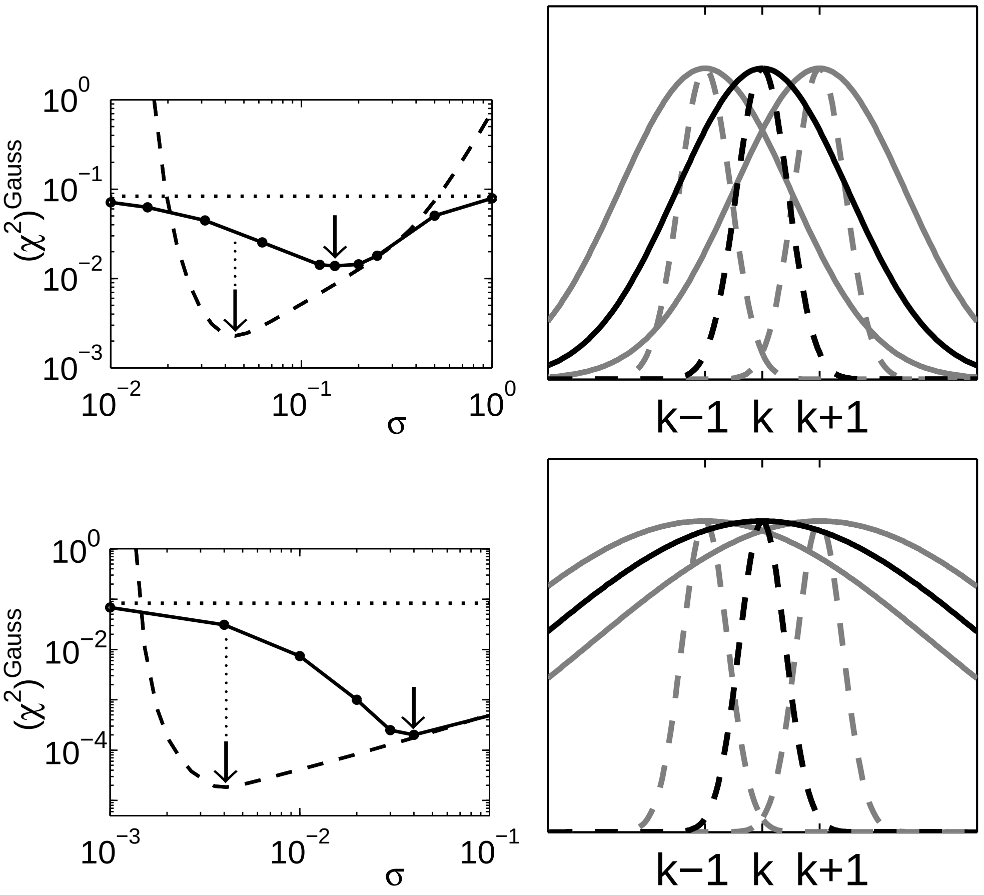

Figure 5.3. Optimal Gaussian tuning. **Left:** Log-log plot of the minimum mean squared error $(\chi^2_{MS})^{Gauss}$ as a function of the scale σ for $f_{max}T = 1$ (solid) compared with the mean asymptotic error $f_{max}T\,(\chi^2_{AS})^{Gauss} = (\chi^2_{AS})^{Gauss} = \mathrm{E}[1/J]$ (dashed) in the case of $N = 10$ (upper) and $N = 100$ neurons (lower). The variance of the *a priori* distribution $\mathrm{Var}[x] = 1/12$ (dotted) provides an upper bound for χ^2_{MS}. The arrows indicate the different minima. **Right:** Comparison of the optimal tuning curves with respect to the mean asymptotic error $(\chi^2_{AS})^{Gauss}$ (dashed) and with respect to the minimum mean squared error $(\chi^2_{AS})^{Gauss}$ (solid) in the case of $N = 10$ (upper) and $N = 100$ neurons (lower). The grey colored lines indicate the adjacent tuning curves.

which implies that the optimal scale with respect to Fisher information is independent of time too. In Fig. 5.3 $f_{max}T\,(\chi^2_{AS}(\sigma))^{Gauss}$ is plotted as a function of the scale in the case of $N = 10$ and $N = 100$ exhibiting a unique minimum, for which the corresponding values are displayed in the following table.

N	σ_{AS}	$f_{max}T\,(\chi^2_{AS})^{Gauss}$
10	0.045	$2 \cdot 10^{-3}$
100	0.004	$2 \cdot 10^{-5}$

As mentioned in section 5.1, the use of the MASE as objective function is justified only in the case $T \to \infty$ of asymptotic normality. For finite T, however, it is necessary to check, whether the MASE agrees with the MMSE or not. Hence, we computed $(\chi^2_{MS})^{Gauss}$ directly for the case of $f_{max}T = 1$ using Monte-Carlo methods (see appendix A.2).

The MMSE as a function of the scale for $f_{max}T = 1$ is also plotted in Fig. 5.3 (left, solid line) and the values of the minima are displayed in the following table.

N	σ_{MS}	$(\chi^2_{MS})^{Gauss}$
10	0.11	$1.2 \cdot 10^{-2}$
100	0.04	$2 \cdot 10^{-4}$

By comparison, we find that the optimal scales with respect to $(\chi^2_{MS})^{Gauss}$ are about one order of magnitude larger than one would conclude from the MASE. In particular, this difference between the short-term optimum and the long-term optimum scale becomes larger, when increasing the number of neurons from ten to hundred. Figure 5.3 (right) shows the corresponding tuning curves illustrating this relative increase. While the MASE is close to $(\chi^2_{MS})^{Gauss}$ for scales that are larger than the optimal scale, the difference between both increases rapidly the more the scale is reduced from the optimal scale and reaches a maximum at the minimum MASE.

5.3 Fisher-optimal codes without tuning curve shape constraints

The analysis of optimal Gaussian tuning in the previous section indicates that Fisher-optimal codes are particularly likely to underestimate the MMSE, if the time window is small. This mismatch between the MASE and the MMSE becomes even more dramatic in the case of Fisher-optimal codes, if one does not stick to the restriction of Gaussian shaped tuning functions. This can be demonstrated by considering Fisher-optimal population codes, where the tuning curves are not subjected to *a priori* constraints apart from a limitation of their dynamic range by a minimum firing rate f_{min} and a maximum firing rate f_{max}. We will first determine the optimal tuning function in the case of a single neuron and then for multiple neurons.

5.3.1 Single neuron

A way to find the tuning function that minimizes the MASE is to start with a calculus of variations for the MASE functional

$$\frac{1}{T}\int\limits_0^1 \frac{f(x)}{(f'(x))^2}\,dx \quad . \tag{5.26}$$

A necessary condition for a minimum of the MASE functional is given by the corresponding Euler-Lagrange differential equation

$$\frac{f(x)}{(f'(x))^2} + 2f'(x)\frac{f(x)}{(f'(x))^3} = C \tag{5.27}$$

which is equivalent to the requirement of a constant Fisher information, because the l.h. side is proportional to $1/J[f]$. The unique solution, satisfying the boundary conditions $f(0) = f_{min}$ and $f(1) = f_{max}$, reads

$$f^{opt}(x) = \left[\left(\sqrt{f_{max}} - \sqrt{f_{min}}\right) x + \sqrt{f_{min}}\right]^2 \quad . \tag{5.28}$$

While the calculus of variation does not account for solutions with kinks, one can prove with some additional effort that f^{opt} in fact constitutes the Fisher-optimal tuning function in the case of the Poisson noise model. Its Fisher information is $J[f^{opt}(x)] = 4T(\sqrt{f_{max}} - \sqrt{f_{min}})^2$. For constant additive Gaussian noise an analog analysis leads to a linear tuning function

$$f(x) = (f_{max} - f_{min})x + f_{min}. \tag{5.29}$$

Since the Fisher information of the minimizer of the MASE, in case of the uniform prior, is a constant, Eq. 5.28 and Eq. 5.29 are also asymptotic minimax.

5.3.2 Many neurons

If x is encoded by more than one neuron, the requirement of identifiability of x does not necessarily imply any more that the tuning functions are monotonic. In particular, if at least one neuron has a strictly monotone tuning curve, all other neurons may have arbitrarily shaped tuning functions. This makes it is easy to construct Fisher-optimal codes, for which the MASE vanishes. In particular, we will show that this is already possible for two neurons only, if they have the following tuning functions (Fig. 5.4):

$$f_1^{wave}(x) = f_{max}x^2\,, \tag{5.30}$$

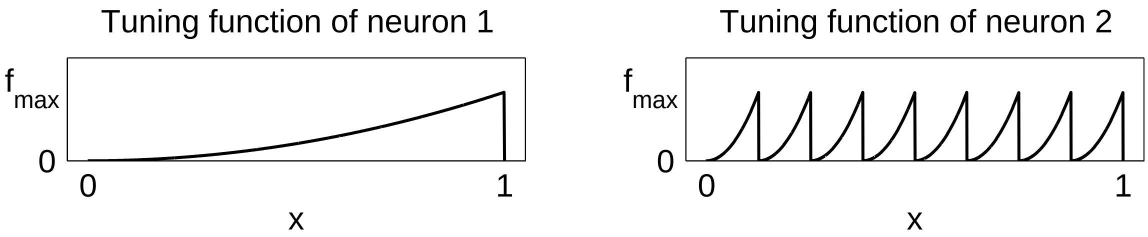

Figure 5.4. Fisher-optimal wave coding scheme consisting of two neurons: One tuning function ensures the identifiability (left). The other tuning function is a wave function that leads to arbitrary large Fisher information when the wave length is decreased (right).

$$f_{2,\omega}^{wave}(x) = f_{max}\ [(\omega x) \bmod 1]^2\ , \tag{5.31}$$

where we have set $f_{min} = 0$ for the sake of clarity. In this example, the total Fisher information is also a constant

$$J[\{f_1^{wave}(x), f_{2,\omega}^{wave}(x)\}] = 4 f_{max} T(\omega^2 + 1)\,. \tag{5.32}$$

Hence, the mean asymptotic error of this wave function encoding scheme equals $(\chi^2_{AS})^{wave} = [4 f_{max} T(\omega^2 + 1)]^{-1}$, which becomes arbitrarily small with increasing ω. If Fisher information would be a general measure for the precision of population codes, this would imply that all coding problems could be solved with two neurons only. However, if we compare Fisher information with the precision of the MS-estimator in the case of $f_{max}T = 1$, we find that $(\chi^2_{MS})^{wave} > 0.06$ for all $\omega \geq 1$.

In summary, this analysis of Fisher-optimal codes allows two important conclusions: (1) With respect to Fisher information, Gaussian tuning curves are particularly disadvantageous, however large or small their tuning widths. (2) Fisher-optimal codes are not necessarily advantageous in case of finite time windows.

5.4 When and why Fisher information fails

In the space of possible arrays of tuning functions it is difficult to name clear cut decision boundaries that tell precisely, where χ^2_{AS} is a fairly good approximation of χ^2_{MS} and where it is not. Therefore, the goal of this section is to gain intuition about which features of an encoding strategy are most relevant to the correspondence between the MASE and the MMSE.

In general, Fisher information and hence the MASE is a separable function in N

and T

$$\chi^2_{AS} = \frac{1}{T} s(N)\,. \tag{5.33}$$

If the *a posteriori* distribution is not normal for small N, but approaches a normal distribution with increasing N, $s(N)$ has to decrease as N^{-1}. This provides a necessary condition for asymptotic efficiency with respect to the limit of large N.

This condition is not necessarily fulfilled, because as becomes clear by the example considered in section 5.3.2, Fisher information can grow arbitrarily fast with increasing N. However, even in case of strongly regularized tuning functions, as in the example of optimal Gaussian tuning investigated above, the population Fisher information grows faster with N than linearly (it grows quadratically), so that in the limit $N \to \infty$, the relationship between Fisher information and the true risk of any estimator through asymptotic normality breaks down for *all* fixed time windows T.

Note, however, that it is indeed possible to construct encodings, for which also the MMSE decreases substantially faster than N^{-1} (an example is given below). This means that the case of asymptotic efficiency holds only for particularly suboptimal codes, which exhibit a high degree of redundancy.

In the example above, there are tuning functions, which map very distant values of x to the same firing rate, and the mismatch between the MASE and the MMSE increases with an increasing number of maxima and minima in the tuning functions. In the following example, we will show that a restriction of the number of maxima is not sufficient to ensure $\chi^2_{AS} \approx \chi^2_{MS}$, but the matching of these two quantities crucially depends on nonlinearities in the tuning functions.

Consider the following class of Fisher-optimal codes built with monotonic tuning functions[7]

$$f^{mono}_{j,\nu}(x) = \begin{cases} f_{max}\left(Nx - \frac{(l-1)N+j-l}{\nu}\right)^2 & , \quad \frac{(l-1)N+j-1}{\nu N} < x < \frac{(l-1)N+j}{\nu N} \\ f_{max}\left(\frac{l}{\nu}\right)^2 & , \quad \frac{(l-1)N+j}{\nu N} < x < \frac{lN+j-1}{\nu N} \end{cases} \,. \tag{5.34}$$

where j denotes the neuron index and $\nu = 1, 2, 3, \ldots$ specifies the shape of the tuning function array (Fig. 5.5). Each tuning curve is completely determined, if one lets l run through all integer values $l = 1, \ldots, \nu$. The Fisher information of these encodings is independent of ν

$$J[\{f^{mono}_{j,\nu}(x)\}_{j=1}^{N}] = 4 f_{max} T\, N^2 \quad . \tag{5.35}$$

In the limit $\nu \to \infty$, this coding scheme can not be distinguished from N identical tuning functions $f^{mono}_{j,\infty}(x) = f_{max} x^2$ (see Fig. 5.5). However, the population Fisher

[7] The proof of Fisher-optimality is based on the same reasoning as the proof of Fisher-optimality for unimodal tuning functions given in section 5.5. It should be mentioned, however, that the set of encodings $\{\{f^{mono}_{j,\nu}(x)\}_{j=1}^{N} : \nu = 1, 2, 3, \ldots\}$ does not contain *all* Fisher-optimal tuning function arrays.

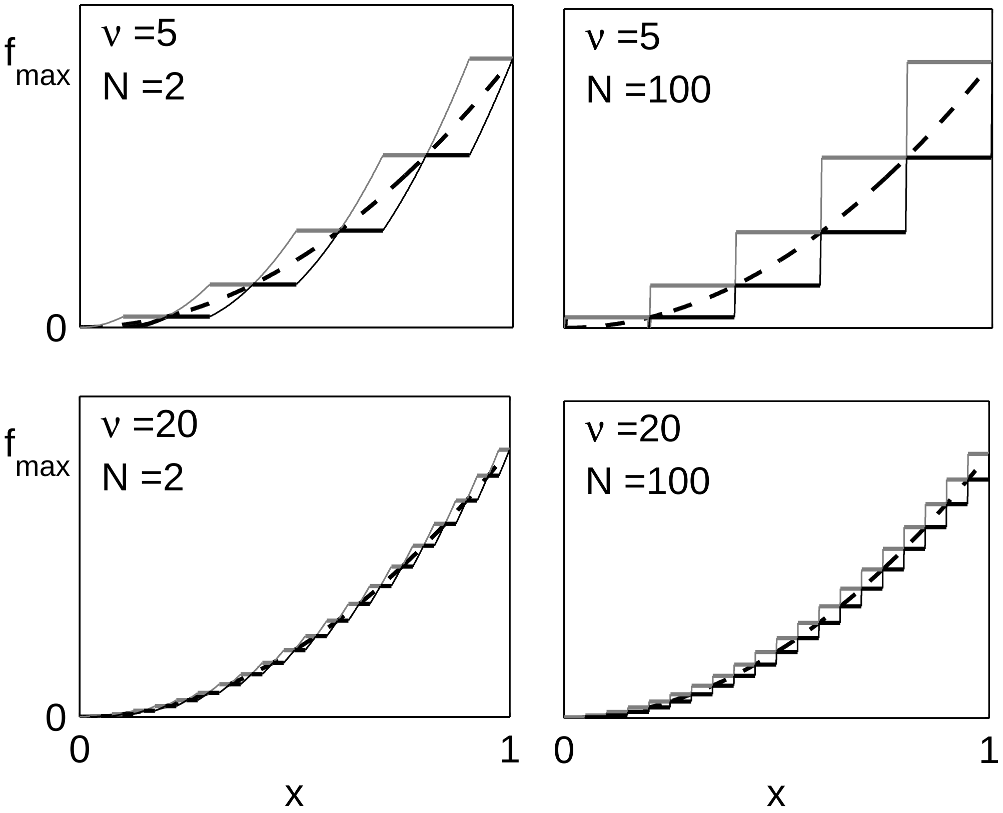

Figure 5.5. Four examples of Fisher-optimal encodings built with monotonic tuning functions as described by Eq. 5.34. The left column shows the case of $N = 2$ and the right column shows the case of $N = 100$, where we only plotted the first tuning function ($j = 1$, grey) and the last one ($j = N$, black). The intermediate tuning functions ($j = 2, \ldots, N - 1$, not shown) are lying in between the first and the last one. Independent of N all tuning functions converge to $f_{\infty}^{mono}(x) = f_{max}x^2$ in the limit $\nu \to \infty$, which is illustrated by the comparison of the case $\nu = 5$ (upper row) with the case $\nu = 20$ (lower row).

information $J[\{f_{j,\nu}^{mono}(x)\}_{j=1}^{N}]$ is N times larger than the population Fisher information of the asymptotic tuning functions $\{f_{j,\infty}^{mono}(x)\}_{j=1}^{N}$ (this is possible, because limiting values are not invariant under a change in the order of limiting processes).

This example demonstrates nicely the fact that Fisher information behaves as if all structures in the tuning functions are of the same relevance independent of their length scale. In fact, however, nonlinearities in the tuning functions become relevant only if they are observable at a scale that is naturally set by the scattering of the noise distribution. Correspondingly, the critical decoding time T_c that is necessary to approach the asymptotic normal case, which is described correctly by Fisher information, increases with increasing ν. For large ν, T_c has to be roughly proportional to ν^2, because the deviations of the tuning functions $f_{j,\nu}^{mono}(x)$ from

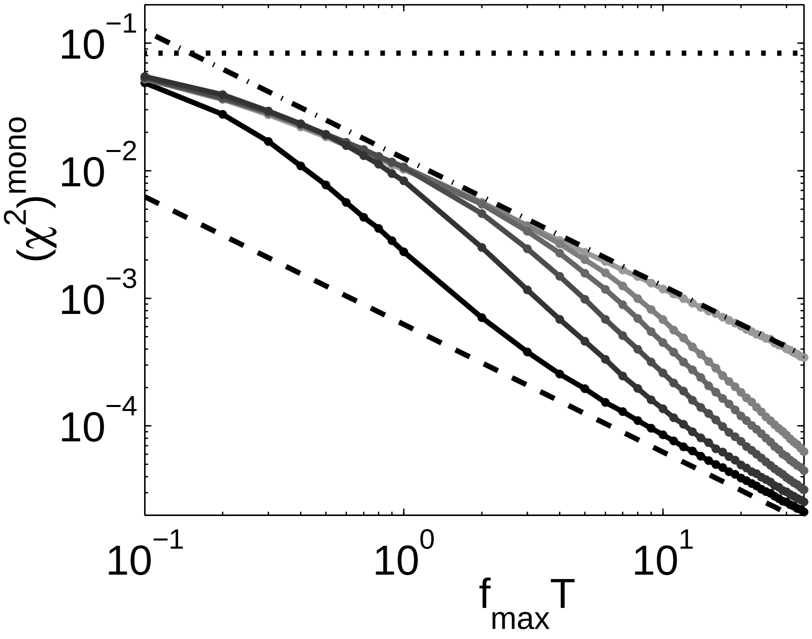

Figure 5.6. The MMSE of $\{f_{j,\nu}^{mono}(x)\}_{j=1}^{20}$ is displayed as a function of the decoding time T for $\nu = 1, 2, 3, 4, 5$ and $\nu = 20$ (solid, from dark to pale). Although all encodings have the same Fisher information (dashed), the critical decoding time increases with increasing ν. If T is smaller than the critical decoding time and larger than $3/(f_{max}N)$, the MMSE curves are well described by the Fisher information of the asymptotic tuning functions $J[\{f_{j,\infty}^{mono}(x)\}_{j=1}^{20}]$ (dot-dashed). For $T < 3/(f_{max}N)$ the bound given by the *a priori* variance 1/12 (dotted) is most relevant.

the 'smoothed' tuning functions $f_{j,\infty}^{mono}(x)$ are of the order of $(1/\nu)^2$ and become relevant, only if the mean squared error, which scales like $1/T$, is of the same order (or smaller). Therefore, the critical decoding time diverges for a diverging ν, which explains the difference in the population Fisher information between $\{f_{j,\nu}^{mono}(x)\}_{j=1}^{N}$ and $\{f_{j,\infty}^{mono}(x)\}_{j=1}^{N}$. Finally, it is worthwhile to note that the ramp coding scheme obtained for $\nu = 1$ can be considered as the best among the class of Fisher-optimal coding schemes built with nondecreasing tuning functions because it has the smallest critical decoding time. This is demonstrated in Fig. 5.6 where it is shown, how the different dependency of $(\chi_{MS}^2)^{mono}$ on the decoding time is effected by the parameter ν.

Taken together, Fisher information is a measure of the long-term coding precision of population codes in the first place, while in the case of finite T one has to check carefully, whether Fisher information provides correct results for the minimum mean squared error. As a rule of thumb, one can say, the smoother the tuning functions, the smaller is the total Fisher information, and the higher the probability that the MASE matches the MMSE.

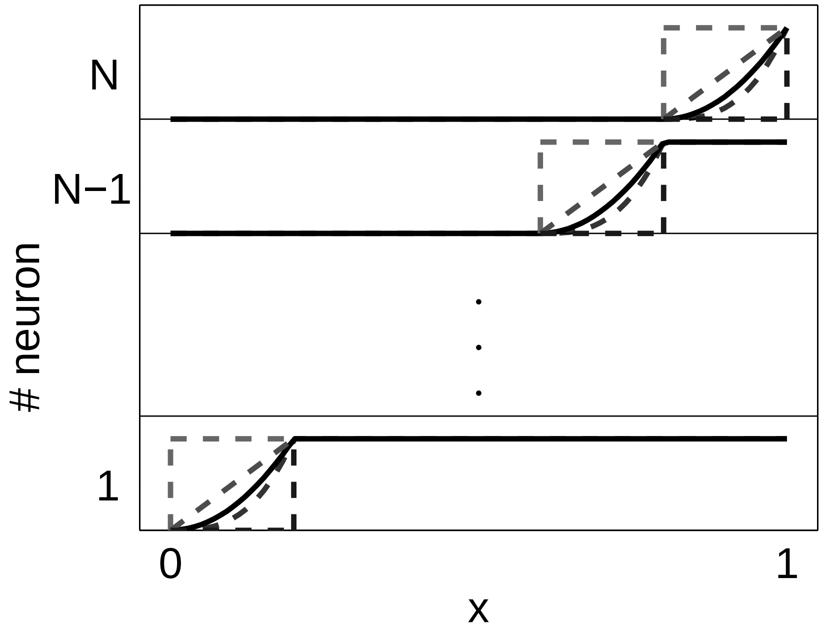

Figure 5.7. Illustration of the ramp coding schemes for different values of α. The tuning functions differ only in the shape of the ramp, which is determined by α. Apart from the Fisher-optimal encoding, which is given for $\alpha = 2$ (solid), we also plotted some other shapes of the ramp (dashed) that correspond to $\alpha = 0, 1, 3, \infty$ (from pale to dark).

Hitherto, we have only considered examples where $\chi^2_{AS} \leq \chi^2_{MS}$, and one might suspect that this holds true in general according to the Cramer-Rao bound. Apart from the trivial fact that χ^2_{AS} diverges in the limit $T \to 0$ in contrast to $\chi^2_{MS} \leq \mathrm{Var}[x]$, it will now be shown that also for arbitrary large T, arrays of tuning functions exist, for which $\chi^2_{AS} >> \chi^2_{MS}$ as well. In order to see this, consider a generalized ramp coding scheme

$$f^{ramp}_{j,\alpha}(x) = f_{max}\left([Nx - j + 1]_+ - [Nx - j]_+\right)^{\alpha}, \tag{5.36}$$

where $[y]_+ = y\Theta(y)$ is the rectifier function. The parameter $\alpha \in [0, \infty)$ can be used to change the tuning curves smoothly from linear ramp functions ($\alpha = 1$) to step functions ($\alpha \to 0$ or $\alpha \to \infty$), which is illustrated in Fig. 5.7. Furthermore, it is important to note that $f^{ramp}_{j,2}(x)$ is identical to $f^{mono}_{j,1}(x)$, which is the Fisher-optimal encoding for nondecreasing tuning functions with the smallest critical decoding time. According to Eq. 5.36 the MASE becomes

$$\left(\chi^2_{AS}(\alpha)\right)^{ramp} = \frac{1}{f_{max}TN^2} \cdot \begin{cases} \frac{1}{\alpha^2(3-\alpha)} & , \ \alpha \in (0,3) \\ \infty & , \ otherwise \end{cases} , \tag{5.37}$$

which implies that for all T, there is an $\alpha < 3$ so that $\left(\chi^2_{AS}(\alpha)\right)^{ramp} >> \mathrm{Var}[x] \geq \left(\chi^2_{MS}(\alpha)\right)^{ramp}$. In the case of $\alpha \geq 3$ the MASE diverges however large T may be.

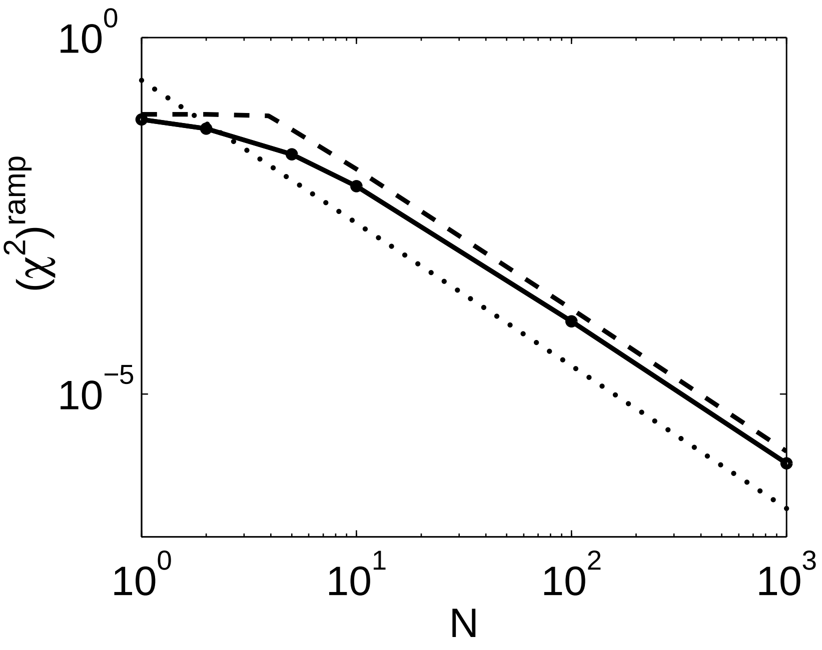

Figure 5.8. The MMSE (solid) of the Fisher-optimal ramp encoding ($\alpha = 2$) is shown in the case of $f_{max}T = 1$. It is very close to the upper bound (dashed) that is the minimum of the two upper bounds that are given by Eq. 5.38 and the *a priori* variance. The dotted line indicates the MASE of the Fisher-optimal ramp encoding, which may be considered as a lower bound on the MMSE for all α provided N is sufficiently large. Therefore all ramp encodings perform similarly well in case of $f_{max}T = 1$.

The strong dependence on α in case of $(\chi^2_{AS}(\alpha))^{ramp}$ is not likely to hold for the MMSE as well. In particular, it is surprising that $(\chi^2_{AS}(3))^{ramp}$ diverges, although the corresponding tuning function array looks very similar to that in the Fisher-optimal case ($\alpha = 2$). The reason for this huge discrepancy is that Fisher information can, in general not account for the precision of encodings, for which $J(x)$ has first-order zeros (or higher-order). One could say that the latter is a weaker form of non-identifiability. Although the encoding is one-to-one, Fisher information cannot account for the precision if the slope of all tuning functions becomes too small somewhere.

In contrast to the strong dependence of the MASE on α, the MMSE in the case of $f_{max}T = 1$ is very similar for all α. It can be shown analytically (see appendix 5.8.2) that the following inequality holds for all α:

$$\left(\chi^2_{MS}\right)^{ramp} \leq \frac{1}{N^2 p} + \frac{1}{N^3}\left(1 - \frac{1-q^N}{p^2}\right) \tag{5.38}$$

where $p = 1 - e^{-f_{max}T}$ increases and $q = 1 - p$ decreases with the length T of the time window. As one can see in Fig. 5.8, this bound is quite close to $(\chi^2_{AS}(2))^{ramp}$ of the Fisher-optimal code.

To summarize, all examples discussed in this section demonstrate that the matching of the MASE with the MMSE depends critically on the effect of nonlinearities of the tuning functions on the stimulus reconstruction. This is also suggested by the fact that Fisher information is calculated only on the basis of the *local* shape of the likelihood function $p(\mathbf{k}|x)$, which corresponds to a linear extrapolation around the true value x_{true}.

In fact, it is possible to give this statement about the 'locality' of Fisher information a precise meaning, because for the most often used noise models like the Poisson noise model considered here or the additive Gaussian noise model, it is actually not necessary to resort to Fisher information, but one can derive the same expressions directly as an approximation of the risk of the MS-estimator via linearization.

If $g_j = f_j^{-1}$ denotes the local inverse function of a tuning curve f_j then the conditional variance $\mathrm{Var}[g_j(k_j/T)|x]$ at the point x can be expressed by

$$\begin{aligned}
\mathrm{Var}[g_j(k_j/T)|x] &= \mathrm{Var}[g_j(f_j(x)) + g_j'(f_j(x))\frac{k_j - Tf_j(x)}{T} + \mathcal{O}(k_j^2)|x] \quad (5.39)\\
&= \left(\frac{g_j'(f_j(x))}{T}\right)^2 \mathrm{Var}\left[k_j|\, x\right] + \mathrm{Var}[\mathcal{O}(k_j^2)|x]\\
&= \left(\frac{1}{Tf_j'(x)}\right)^2 Tf_j(x) + \mathrm{Var}[\mathcal{O}(k_j^2)|x]\\
&= \frac{f_j(x)}{Tf_j'^2(x)} + \mathrm{Var}[\mathcal{O}(k_j^2)|x] \quad (5.40)\\
&= \frac{1}{J[f_j(x)]} + \mathrm{Var}[\mathcal{O}(k_j^2)|x]\,. \quad (5.41)
\end{aligned}$$

In a similar way, Fisher information shows up if one determines the error of the MS-estimator in the limit of vanishing noise. For any given x the MS-estimator can be approximated by a linear function of $\mathbf{k}$ in this limit. In particular, this linear function can be set to the form of a superposition of the inverse tuning functions $g_j = f_j^{-1}$, because the MS-estimator is asymptotically unbiased. Therefore, it holds

$$\hat{x}_{MS}(\mathbf{k}) \approx x + \sum_j W_j g_j'(f(x))\,(k_j/T - f_j(x)) = x + \sum_j W_j \frac{k_j/T - f_j(x)}{f_j'(x)}\,, \quad (5.42)$$

where the $\{W_j\}_{j=1}^N$ stand for an arbitrary weighting with $\sum_{j=1}^N W_j = 1$. Accordingly, the conditional error variance of the MS-estimator is given by

$$\mathrm{E}[(\hat{x}_{MS}(\mathbf{k}) - x)^2|x] = \sum_j W_j^2 \frac{\mathrm{Var}[k_j|x]}{(Tf_j'(x))^2}\,. \quad (5.43)$$

Minimizing Eq. 5.43 under the constraint of $\{W_j\}_{j=1}^N$ yields

$$W_j = \frac{(Tf_j'(x))^2}{\mathrm{Var}[k_j|x] \sum_j \frac{(Tf_j'(x))^2}{\mathrm{Var}[k_j|x]}} \tag{5.44}$$

and correspondingly the conditional error variance becomes

$$\mathrm{E}[(\hat{x}_{MS}(\mathbf{k}) - x)^2|x] = \frac{1}{\sum_j \frac{(Tf_j'(x))^2}{\mathrm{Var}[k_j|x]}} = \frac{1}{J[\{f_j(x)\}_{j=1}^N]}\,. \tag{5.45}$$

This somewhat heuristic calculation suggests that the inverse Fisher information is a good approximation of the risk of the MS-estimator, when the scattering of the MS-estimator around the true value of x is restricted to a region within which the tuning functions may be considered as linear (see also [Kay93]).

5.5 Optimizing unimodal tuning functions

The optimization of the width of unimodal tuning functions, which has been discussed in the previous chapter 4, is built upon the assumption of a given, fixed tuning profile. In contrast, here it shall be investigated, what the Fisher-optimal encoding looks like, if one allows for adjustment of the profile of the tuning functions as well. While we have seen that Fisher information can diverge already in the case of only two neurons if no particular constraints are imposed on the shape of the tuning function, the MASE remains finite if the number of maxima of each tuning function is set to be limited. This is clearly the case for unimodal tuning functions, which have one maximum only. For such encodings, Fisher information cannot increase faster than proportional to N^2, provided identifiability of x.

In order to avoid asymmetries due to the boundaries of the interval, we now switch to the case of a circular random variable (which could represent e.g. the angle of an oriented bar). The ring topology of the circular random variable requires to modify the Euclidean distance slightly

$$D(x_1, x_2) = \min\{|x_1 - x_2|, 1 - |x_1 - x_2|] \tag{5.46}$$

by always choosing the path of smaller distance. Accordingly the MMSE is then given by

$$\left(\chi_{MS}^2\right)^{uni} = \mathrm{E}[D(\hat{x}_{MS} - x)^2]\,. \tag{5.47}$$

While this modification reduces the *a priori* error, compared to the case without periodic boundary conditions, it has no effect asymptotically.

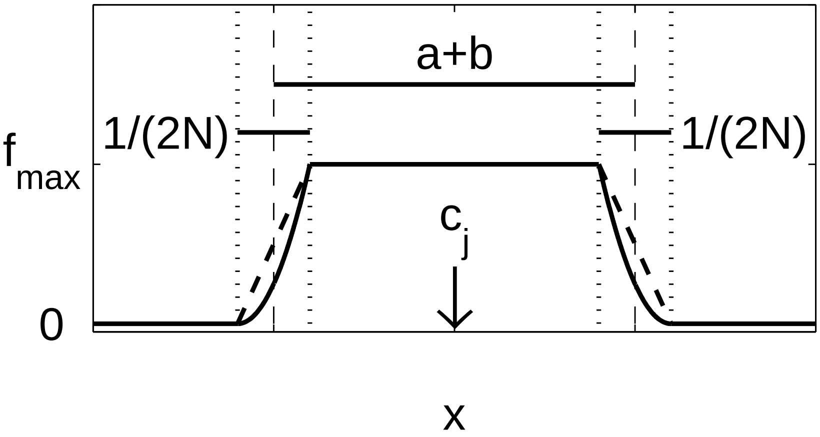

Figure 5.9. The Fisher-optimal unimodal tuning curve with the smallest critical decoding time is flat with small edges of length $1/(2N)$. Fisher information and the special type of noise model is relevant for the shape of optimal tuning curves, only within these edge regions: The solid line refers to a Poisson noise model and the dashed line to additive Gaussian noise of arbitrary variance.

For convenience, assume that x is encoded by unimodal symmetric tuning curves of identical shape with equidistantly distributed centers $c_j = j/N$

$$f_j^{uni}(x) = \begin{cases} f_{max} & , \quad D(x,c_j) \leq a \\ g\left(\frac{D(x,c_j)-a}{b-a}\right) & , \quad a < D(x,c_j) < b \\ f_{min} & , \quad b \leq D(x,c_j) \leq \frac{1}{2} \end{cases} \quad , \tag{5.48}$$

where $g : (0,1) \to [f_{min}, f_{max}]$ is a monotone decreasing and otherwise arbitrary function.

Since the Fisher information $J[f_j^{uni}(x)]$ can be positive only for $a < D(x,c_j) < b$, it is refered to the corresponding regions as *Fisher information regions (F-regions)* of the tuning functions. Independent of the function $g(z)$, $J[f_j^{uni}(x)]$ is proportional to the inverse squared F-region width $(b-a)^{-2}$. Therefore, it is a necessary condition for a minimum of the MASE, that the F-regions of different neurons must not overlap, because the contributions of different neurons add at most linearly to the total Fisher information.

Then the evaluation of the MASE can be decomposed

$$\mathrm{E}\left[\frac{1}{J^{uni}(x)}\right] = \int\limits_{D(x,c_j)<a} \frac{dx}{J^{uni}(x)} + \int\limits_{a<D(x,c_j)<b} \frac{dx}{J[f_j^{uni}(x)]} + \int\limits_{b<D(x,c_j)<0.5} \frac{dx}{J^{uni}(x)} \tag{5.49}$$

and we can conclude from chapter 4.1. that $g(z) = ((\sqrt{f_{max}} - \sqrt{f_{min}})(1-z) + \sqrt{f_{min}})^2$ is the minimizer of the second term. It remains to determine the optimal choice of

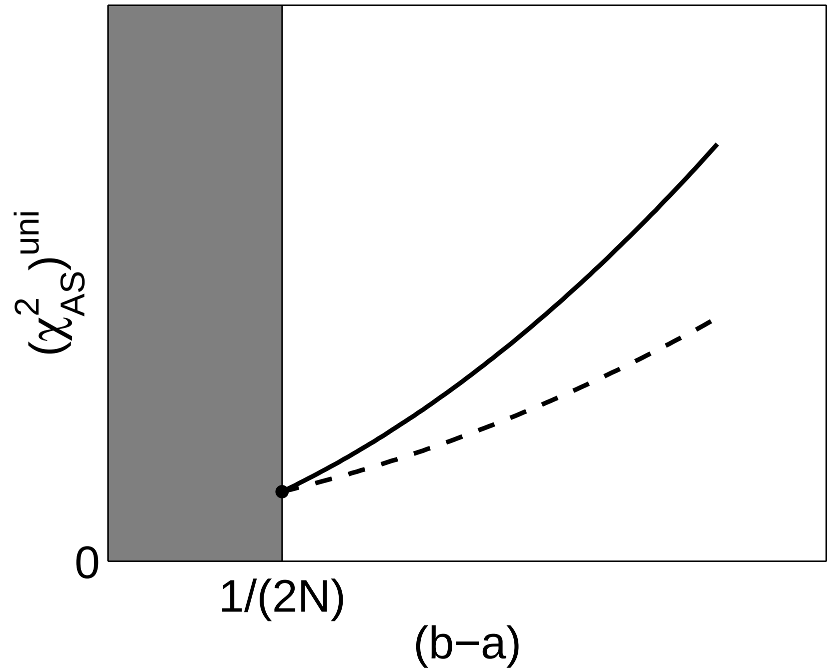

Figure 5.10. Sketch of the MASE (solid) as a function of the F-region width $(b-a)$. If $(b-a) < 1/(2N)$ there is an interval with zero Fisher information and hence the MASE diverges (grey region). For $(b-a) = 1/(2N)$ we have derived the encoding with minimal MASE, for which we found that it equals $1/\mathrm{E}[J]$ of that encoding (dot). Since the MASE is larger or equal to $1/\mathrm{E}[J]$ (dashed) and this lower bound is an increasing function of $(b-a)$, the MASE is an increasing function of $(b-a)$ too, independent of the particular shape of the encoding. It follows that the encoding that minimizes the MASE in case of $(b-a) = 1/(2N)$ is also Fisher-optimal, compared to coding schemes with different F-region widths.

a and b. As we will show in the following, the MASE becomes a minimum, if the F-region width is set to $b-a = 1/(2N)$, and $b = k/N$ for any $k \in \{1, 2, \ldots, N\}$ (Fig. 5.9). In this case, the total Fisher information J^{uni} does not depend on x so that it holds

$$\mathrm{E}\left[\frac{1}{J^{uni}}\right] = \frac{1}{\underset{x}{\mathrm{E}}[J^{uni}]} = \frac{1}{16\left(\sqrt{f_{max}} - \sqrt{f_{min}}\right)^2 TN^2}\,. \tag{5.50}$$

Because $\mathrm{E}[J^{uni}]$ decreases if $b-a$ is increased, it follows together with Jensen's inequality

$$\mathrm{E}\left[\frac{1}{J}\right] \geq \frac{1}{\underset{x}{\mathrm{E}}[J]} \tag{5.51}$$

that the optimal F-region width $b-a$ cannot be larger than $1/(2N)$ (see Fig. 5.10). However, $b-a$ can not be smaller than $1/(2N)$ either, due to the requirement of identifiability as well as due to the fact that the MASE diverges for all encoding strategies with $b - a < 1/(2N)$.

Since the identical minimal MASE is achieved for sharp tuning as well as for broad tuning, we recovered the result of the scaling rule (4.9) in case of $D = 1$ that the length of the tuning width $w := a + b$ is not a characteristic signature of Fisher-optimal codes. Instead of the tuning width, we find that in general, the length of the F-region width is crucial for Fisher-optimality, because the optimization is mainly a matter of making the F-region width as small as possible (cf. Fig. 5.10) as has been demonstrated by the example of unimodal tuning functions. Accordingly, steep changes and flat plateaus are the striking signatures of Fisher-optimal tuning curves, which is the more true, the larger the populations are, because the minimal average F-region width scales at least as $1/N$.

While the analysis above suggests that Fisher-optimal unimodal tuning functions are approximately box-shaped if N is sufficiently large, it is important to note that the set derived above is not complete, because we have imposed various additional constraints on the tuning functions there. If we drop these assumptions, there are actually many more Fisher-optimal unimodal tuning functions, all having the same MASE as given by Eq. 5.50. E.g. if we require monotony instead of *strict* monotony for $g(z)$ only, g may have arbitrarily many constant parts. Then similar to the idea underlying the Fisher-optimal class of monotonic tuning function encodings given by (5.34), this allows to construct various Fisher-optimal codes, which have no features in common apart from the strongly bimodal distributions of their derivatives.

5.6 Optimal tuning width - a question of energy?

While it was not possible to determine an optimal tuning width with respect to Fisher information under the constraints considered above, it is suspected that energy consumption constitutes an important design constraint for neuronal processing in the brain [LB96], and as pointed out by several researchers, sparse codes are advantageous under this conditions [Bad96]. In population coding models the (life-time[8]) sparseness of a neuronal encoding is given by the mean of the tuning function, which is the firing rate of the neuron averaged over the prior distribution $p(x)$. In case of a flat prior distribution and unimodal tuning, the mean of a tuning function is essentially determined by the tuning width. Therefore, we can suspect to find a symmetry breaking towards small tuning widths, if a limit on energy consumption is taken into account.

[8] Life-time sparseness refers to the neuronal activity averaged over time, which is distinguished from population sparseness, which refers to the average activity of a population at given instant of time. Population sparseness implies life-time sparseness, but the opposite is not true.

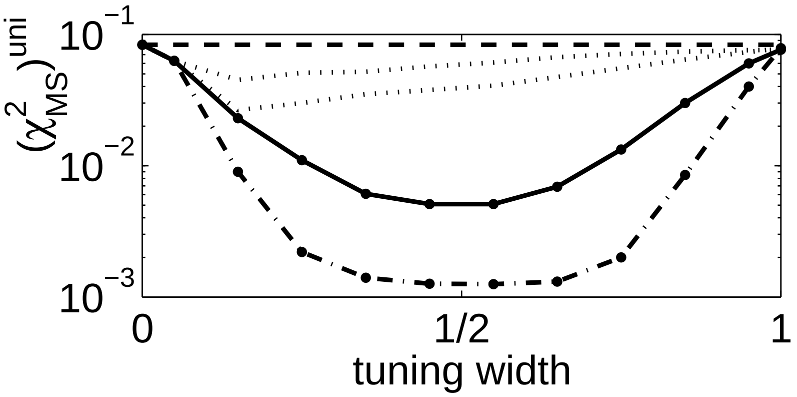

Figure 5.11. MMSE as a function of the tuning width in the case of $N = 10$ neurons (solid) and $N = 20$ neurons (dot dashed). The dashed line indicates the *a priori* variance $\mathrm{Var}[x] = 1/12$ that is an upper bound for χ^2_{MS}. The dotted lines denote the results for $N = 10$ (upper) and $N = 20$ (lower) neurons if the energy constraint (Eq. 5.53) is taken into account.

Hitherto, the coding efficiency was only limited by a constraint on the power ($f(x) \leq f_{max}$), which can be motivated for instance by the refractory period of a neuron after the generation of an action potential. On the other hand, however, it is likely that energy constraints are relevant, because the average interspike intervals of cortical neurons are much larger than their refractory period. This fact can be taken into account, if one assumes an additional upper bound for the mean firing rates $\mathrm{E}[f_j(x)] \leq \langle f \rangle_{max}$.

This constraint can be applied to the Fisher-optimal tuning functions derived above by computing their mean value as a function of the tuning width:

$$\mathrm{E}\left[f_j^{optuni}(x)\right] = \left(w - \frac{1}{6N}\right) f_{\max} \tag{5.52}$$

Then, both constraints together can be expressed by a width dependent maximum firing rate

$$\tilde{f}_{max}(w) = \min\left(f_{max}, \frac{\langle f \rangle_{\max}}{w - \frac{1}{6N}}\right) . \tag{5.53}$$

It is straightforward to see that under this constraint the MASE is minimal only for those tuning widths that are smaller than or equal to $\langle f \rangle_{\max} + 1/(6N)$.

Since the inverse Fisher information will not be close to the MMSE in case of particularly small (and broad) tuning, we computed the MMSE numerically for $f_{min} = 0$ and $f_{max}T = 1$ (this corresponds e.g. to $f_{max} = 200$ Hz and $T = 5$ ms) in the case of $N = 10$ and $N = 20$ neurons.

In the absence of energy constraints (i.e. $\langle f \rangle_{\max} \geq f_{\max}$), it turns out that $w \approx 1/2$

is optimal, while the objective function is flat in a wide region around this optimum (Fig.5.11). There is a slight asymmetry due to the rate-dependent noise of the Poisson model, such that $(\chi_{MS}(w))^{uni}$ increases not so fast for decreasing tuning width than for increasing width. However, this asymmetry is rather negligible.

While the broadness of the minimum of the MMSE, as a function of the tuning width, does not indicate a substantial advantage of a certain receptive field size, this changes if energy consumption is taken into account. In order to demonstrate the effect of this energy constraint, the MMSE has been determined numerically in the case, where the mean firing rate is limited by $\langle f \rangle_{max} = f_{max}/20$ (this e.g. corresponds to a maximum firing rate of 200 Hz and a mean firing rate of 10 Hz). The resulting $(\chi^2_{MS})^{uni}_{energy}$ is shown in the case of $f_{max}T = 1$ in Fig. 5.11 by the dotted line exhibiting a distinct minimum located directly at the energy bound.

This preference for smaller tuning widths gives a precise meaning to the statement that sparse coding can be explained by constrained energy consumption. In contrast to previous conclusions in [vV01], this result does not rely critically on the Poisson noise model, but holds also true in case of a Gaussian noise model. In fact, the Fisher-optimal code for the Gaussian noise model differs from the Poisson case only by a small change of the function $g(z)$, which becomes $g(z) = (f_{max} - f_{min})(1 - z) + f_{min}$ (see Fig. 5.9, dashed line). This leads to a MASE $v/[2N(f_{max} - f_{min})]^2$, where $v = \mathrm{Var}[k_j|x]$ denotes the constant noise variance. Correspondingly, the MASE is again minimal for $w \leq \langle f \rangle_{max}/f_{max} \leq 1/20$ in the same way as in the Poisson case. Moreover, the Poisson model itself is not sufficient to explain small receptive fields, because $(\chi^2_{AS})^{uni}$ is identical for all tuning widths $w = 1/N, 2/N, \ldots, 1$ and the asymmetry of $(\chi^2_{MS})^{uni}$ in the absence of energy constraints is very weak.

5.7 Discussion and conclusion

The goal of this work was to understand the general principles of optimal population coding constraining the possibilities of *inter-neuronal* signal processing in cortex. To this end, Fisher-optimal encodings have been derived within much larger sets of candidate tuning function arrays than in previous studies, because the usual restriction to bell-shaped tuning functions cannot be justified in the context of efficient coding. In particular, the ultimate Fisher-optimal encoding has been determined by a nonparametric calculus of variation. In case of a single neuron, this leads to a quadratically increasing tuning function as the unique optimum. In case of two neurons, there are already infinitely many possibilities to achieve a one-to-one encoding with a uniformly diverging Fisher information.

The latter is clearly an example, which demonstrates that Fisher information is not always appropriate as a measure of coding accuracy and hence raises the question what one can actually learn from the comparison of encoding strategies with respect to Fisher information. Although Fisher information is predominantly used in the population coding literature, there is hardly any mention about its limited range of validity. Since it is indispensable to have at least an intuition about the conditions under which an approximation is likely to fail, both, the mathematical background as well as several instructive examples have been presented.

In particular, it has been pointed out that a necessary condition for asymptotic efficiency is that the population Fisher information does not increase faster with N than linearly, which holds only for particularly suboptimal codes. Consequently, even for the graceful example of Gaussian tuning curves, the use of Fisher information is awkward, if the scale of the Gaussians is not fixed but decreased with increasing N (as it was advantageous). For illustration of this fact, the Fisher-optimal scale of Gaussian tuning curves has been compared with the MMSE-optimal width in case of $N = 10$ and $N = 100$ neurons. While the Fisher-optimal scale decreased by a factor of ten, the MMSE-optimum decreased only by a factor of smaller than three. In both cases ($N = 10$ and $N = 100$), the Fisher-optimal scale was much smaller than the MMSE-optimum.

Furthermore, this example illustrates that the shape of an optimal encoding depends on the available decoding time, while the Fisher-optimal codes are always independent of T. This does not come as a surprise, if one accepts the interpretation that the Fisher information divided by T reflects the *rate* with which the error decreases in the limit $T \to \infty$. From this point of view, the crucial question is, how large does T have to be for a given tuning function array, at least so that Fisher information describes the risk sufficiently correct.

As a partial answer to this question, it has been demonstrated that the critical decoding time T_c that is necessary for a sufficient matching of the MASE and the MMSE, is typically increased by nonlinearities of the tuning functions. In particular, T_c grows with the frequency with which the tuning functions rise and decay between their minimum and maximum firing rate.

The relevant counting time window length can be related to the time scale at which neurons integrate over their synaptic inputs. Since the high degree of irregularity of neuronal discharge in cortex [SK93] implies that the effective integration time constant is of the order of a few milliseconds, the situation, where the spike count of a single neuron is of the order of one, appears to be most relevant. Clearly, the MASE cannot be expected to be a reliable measure for the MMSE in that case. Nevertheless, it might give worthwhile hints about the qualitative behavior of the MMSE that can be used to guide numerical studies of the MMSE, which typically require substantial computer power.

If Fisher information is used as an objective function in order to determine optimal coding schemes, one is typically lead to tuning functions with a slope that is as large as possible. In fact, the Fisher information of a (bounded) tuning function behaves like a penalty term for regularization. Because Fisher information is intended to become as large as possible, however, Fisher-optimality has quite the opposite effect of regularization and hence, it cannot be expected to rule out a large number of coding schemes on the basis of Fisher information only.

If the total number of maxima of the tuning functions is finite, the MASE remains finite, too. However, also in these cases, there is no unique Fisher-optimal code, but very many encodings achieve the same minimal MASE. This can be conceived e.g. from the case of nondecreasing tuning functions, for which we presented an infinitely large set of Fisher-optimal encodings (that was still not complete). The method with which Fisher optimal codes can be derived was presented in the case of unimodal tuning functions. It is, however, not restricted to monotonic or unimodal tuning functions, but can also be applied to tuning functions with more maxima. The crucial point is that the F-regions of Fisher-optimal tuning functions do not overlap and the total length of the F-regions of a single tuning function f_j equals $d_j / \sum_{i=1}^{N} d_i$, where d_j denotes the number of how many times $f_j(x)$ is allowed to traverse the dynamic range[9]. In the most simple case, the tuning function f_j then has d_j regions, within which the tuning function increases (or decreases) quadratically as given by Eq. 5.28. We therefore have the general formula for the Fisher information of Fisher-optimal codes:

$$J = 4T \left(\sqrt{f_{max}} - \sqrt{f_{min}} \right)^2 \left(\sum_{j=1}^{N} d_j \right)^2 . \tag{5.54}$$

It is, however, also possible, that the quadratic increase itself is interrupted by flat regions as it is the case e.g. for $f_{j,\nu}^{mono}(x)$ and $\nu > 1$, so that the F-regions can be shattered over the entire range of x. Due to this freedom, the number of Fisher-optimal codes is very large.

Furthermore, we found that the question whether small or broad tuning widths are advantageous cannot be decided, if the number of neurons N, the available decoding time T and the maximum firing rate f_{max} are given only. Instead, it was shown that a limitation of the average firing rate, which can be motivated by energy consumption, naturally breaks the symmetry towards sparse codes with small tuning widths.

While the tuning width cannot serve as a general indicator of Fisher-optimality, the minimization of the dynamic range is the foremost property of such. Fisher-optimal codes, however, do not perform equally well, with respect to the MMSE. The example of the Fisher-optimal monotonic tuning curves demonstrates a clear advantage for those tuning functions which are mostly binary ($\nu = 1$). This suggests that a strong selectivity, that is when the cells switch rapidly between their minimum and maximum firing rates so that they have a small dynamic range only, appears to

[9] E.g. $d_j = 1$ in case of monotonic tuning functions and $d_j = 2$ in case of unimodal tuning functions.

be a very general signature of efficient encodings. This holds not only asymptotically for long decoding time windows, but even for short windows as well. In order to further test this hypothesis, the MMSE is computed and compared for a selection of characteristic encoding strategies, which is the subject of the next chapter.

5.8 Appendix

5.8.1 Fisher information and the exponential family

The mean squared error of any estimator can always be decomposed into its bias $b_{\hat{x}}(x) = (g(x) - x)^2$ and its variance $v_{\hat{x}}(x) = \mathrm{E}[(\hat{x} - g(x))^2|x]$, where we introduced $g(x) = \mathrm{E}[\hat{x}|x]$ for the sake of clarity. Accordingly, the Cramer-Rao bound (Eq. 5.10) can also be given in the form

$$v_{\hat{x}}(x) \geq \frac{g'(x)^2}{J(x)} \,. \tag{5.55}$$

For this lower bound it is known that equality holds, if and only if $p(\mathbf{k}|x)$ constitutes an exponential family. While the Poisson distribution constitutes an exponential family with respect to the mean spike count $\mu_j = f_j(x)T$ for each neuron j, it depends on the shape of the tuning functions, whether an estimator of x exists, for which equality holds in Eq. 5.55. Such an estimator has to satisfy the following equation (see e.g. Lehmann & Casella, 1999)

$$\hat{x}(\mathbf{k}) = g(x) + \frac{g'(x)}{J(x)} \partial_x \log p(\mathbf{k}|x) \,. \tag{5.56}$$

Therefore the mean squared error of an estimator is completely determined by Fisher information, if it is unbiased (i.e. $g(x) = x$) and the r.h. side of Eq. 5.56 is independent from x, i.e.

$$x + \frac{\partial_x \log p(\mathbf{k}|x)}{J(x)} = const \,. \tag{5.57}$$

Inserting Eq. 5.1 and taking the derivative with respect to x yields

$$\frac{(k - \mu)\mu''}{\mu'^2} = 0 \tag{5.58}$$

in case of $N = 1$. From this it follows that μ'' equals zero and hence, the tuning function $f(x) = \frac{1}{T}\mu(x)$ is required to be linear. While we didn't solve Eq. 5.57 for $N > 1$, this short calculation may hint towards the strong restrictions that it imposes on the shape of the tuning functions.

5.8.2 Derivation of equation 5.38

The upper bound results from a calculation of the error of an suboptimal estimator:

$$\hat{x}(\mathbf{k}) := \max\left\{\frac{1}{N}, \frac{j}{N}\Theta\left(k_j - \frac{1}{2}\right)\middle| j = 1,\dots,N\right\} \tag{5.59}$$

We decompose the mean squared error

$$\begin{aligned}\left(\chi^2(\alpha)\right)^{ramp} &= \int\limits_0^1 \sum_{\mathbf{k}} (x-\hat{x}(\mathbf{k}))^2 p(\mathbf{k}|x)dx \\ &= \frac{1}{N}\sum_{a=1}^{N} \underbrace{N \int\limits_{\frac{a-1}{N}}^{\frac{a}{N}} \sum_{\mathbf{k}} (x-\hat{x}(\mathbf{k}))^2 p(\mathbf{k}|x)dx}_{\chi_a^2} \qquad (5.60)\end{aligned}$$

and consider the parts χ_a^2, $a = 1,\dots,N$ separately.

$$\chi_1^2 = N\int\limits_0^{\frac{1}{N}}\left(x-\frac{1}{N}\right)^2 dx \le \max_{x\in[0,1/N]}\left(x-\frac{1}{N}\right)^2 = \frac{1}{N^2} \tag{5.61}$$

$$\begin{aligned}\chi_2^2 &= p(k_2>0|x\in[1/N,2/N])\int\limits_{\frac{1}{N}}^{\frac{2}{N}}\left(x-\frac{2}{N}\right)^2 dx \\ &\quad +p(k_2=0|x\in[1/N,2/N])\int\limits_{\frac{1}{N}}^{\frac{2}{N}}\left(x-\frac{1}{N}\right)^2 dx \\ &\le p(k_2>0|x\in[1/N,2/N])\max_{x\in[1/N,2/N]}\left(x-\frac{2}{N}\right)^2 \\ &\quad +p(k_2=0|x\in[1/N,2/N])\max_{x\in[1/N,2/N]}\left(x-\frac{1}{N}\right)^2 \\ &\le p(k_2>0|x\in[1/N,2/N])\frac{1}{N^2}+p(k_2=0|x\in[1/N,2/N])\frac{1}{N^2} \\ &= \frac{1}{N^2} \qquad (5.62)\end{aligned}$$

$$\begin{aligned}\chi_3^2 &\le p(k_2>0|x\in[2/N,3/N])\int\limits_{\frac{2}{N}}^{\frac{3}{N}}\left(x-\frac{2}{N}\right)^2 dx + p(k_2=0|x\in[2/N,3/N])\left(\chi_2^2+\frac{1}{N^2}\right) \\ &\le \frac{1}{N^2}(1+p(k_2=0|x\in[2/N,3/N])) \qquad (5.63)\end{aligned}$$

In general it holds for all $a \geq 2$:

$$\chi^2_{a+1} \leq \frac{1}{N^2} + p(k_a = 0 | x \in [a/N, (a+1)/N])\chi^2_a = \frac{1}{N^2} + q\,\chi^2_a\,, \tag{5.64}$$

where we introduced the abbreviation q for the probability $p(k_a = 0|x \in [a/N, (a+1)/N]) = e^{-f_{max}T}$, which does not depend on the neuron index. Together with $p := 1 - q$ it then follows by induction:

$$\chi^2_a \leq \frac{1}{N^2}\,\frac{1-q^{a-1}}{1-q} = \frac{1}{N^2}\frac{1-q^{a-1}}{p}\,, \tag{5.65}$$

because it holds

$$\chi^2_{a+1} \overset{Eq.\ 5.64}{\leq} \frac{1}{N^2} + q\,\frac{1}{N^2}\frac{1-q^{a-1}}{p} = \frac{1}{N^2}\left(\frac{p+q-q^a}{p}\right) = \frac{1}{N^2}\,\frac{1-q^a}{p}\,. \tag{5.66}$$

According to (5.60) we finally obtain

$$\begin{aligned}
\left(\chi^2(\alpha)\right)^{ramp} = \frac{1}{N}\sum_{a=1}^{N}\chi^2_a \quad &\leq \quad \frac{1}{N}\left(\frac{1}{N^2} + \sum_{a=2}^{N}\frac{1}{N^2}\,\frac{1-q^{a-1}}{p}\right) \\
&= \quad \frac{1}{N^3}\left(1 + \frac{1}{p}\sum_{a=2}^{N}(1-q^{a-1})\right) \\
&= \quad \frac{1}{N^3}\left(1 + \frac{N-1}{p} - \frac{1}{p}\sum_{a=0}^{N-2}q^{a+1}\right) \\
&= \quad \frac{1}{N^3}\left(1 + \frac{N-1}{p} - \frac{q}{p}\frac{1-q^{N-1}}{1-q}\right) \\
&= \quad \frac{1}{N^2 p} + \frac{1}{N^3}\left(1 - \frac{p+q-q^N}{p^2}\right) \\
&= \quad \frac{1}{N^2 p} + \frac{1}{N^3}\left(1 - \frac{1-q^N}{p^2}\right)\,.
\end{aligned} \tag{5.67}$$

Chapter 6

MMSE-Optimal Codes: Labels or Intensity?

In principle, one can distinguish two extreme cases of population encoding strategies: in case of *label-pattern coding*[1] the neurons function as binary devices, which are either (maximally) active or inactive (see Fig. 6.1 right). In the other extreme of *intensity coding*, the response of each neuron constitutes a (noisy) representation of an analog value, which can be reliably read out by averaging or pooling over the responses of all neurons (see Fig. 6.1 left).

In the context of large-scale neural network modeling and studies of neuronal population dynamics, the idea of population coding plays an important role as well. There, populations of neurons are commonly taken as the basic units of information processing, computing in their totality a single analog number. While averaging over sufficiently large populations of neurons clearly allows for precise analog rate estimates at short time scales (see Fig. 6.2), the results shown in the previous chapter suggest that optimal encodings of an analog signal might be closer to label-pattern coding than to intensity coding. Because of the limited validity of Fisher information in particular for tuning functions with a small dynamic range, we now study how relevant intensity coding is for efficient analog population signaling if the MS-estimator is taken as neuronal read out.

To this end, we consider examples from two different classes $\mathcal{R}_2^N, \mathcal{W}^N$ of tuning function arrays, where N indicates the total number of tuning functions per array. A tuning function array in $\mathcal{R}_2^N$ is of the form $\{f_j^{\lambda,N-ramp}\}_{j=1}^N$, where each tuning

[1] I introduce the new term 'label-pattern' coding, because the related terms 'place coding' and 'labeled-line' coding have been used by several authors in order to refer to the special case, where only one neuron of a population is active at a time.

curve is a parabolic ramp function

$$f_j^{\lambda,N-ramp}(x) = \begin{cases} 0 & , \quad 0 \le x \le X_{j,\lambda,N} \\ f_{max}\left(\frac{x - X_{j,\lambda,N}}{\lambda}\right)^2 & , \quad X_{j,\lambda,N} < x < X_{j,\lambda,N} + \lambda \\ f_{max} & , \quad X_{j,\lambda,N} + \lambda \le x \le 1 \end{cases} . \tag{6.1}$$

In Eq. 6.1 it has been used the abbreviation

$$X_{j,\lambda,N} = \frac{j-1}{N-1}(1-\lambda) \tag{6.2}$$

and λ determines the length of the region within which the ramp function is increasing (Fig. 6.3).

The tuning function arrays in $\mathcal{W}^N$ are arbitrary combinations $(\nu_1, \ldots, \nu_N)$ of square wave functions

$$f^{\nu-wave}(x) = f_{max}\Theta((2^{\nu-1}x) \bmod 1 \; - 0.5) \tag{6.3}$$

with $\nu \in \{1, 2, 3, \ldots\}$.

As in the previous chapter, we again assume a factorizing Poisson noise model

$$p(\mathbf{k}|x) = \prod_{j=1}^{N} p(k_j|\mu_j(x)) = \prod_{j=1}^{N} \frac{(\mu_j(x))^{k_j}}{k_j!} \exp\{-\mu_j(x)\} \tag{6.4}$$

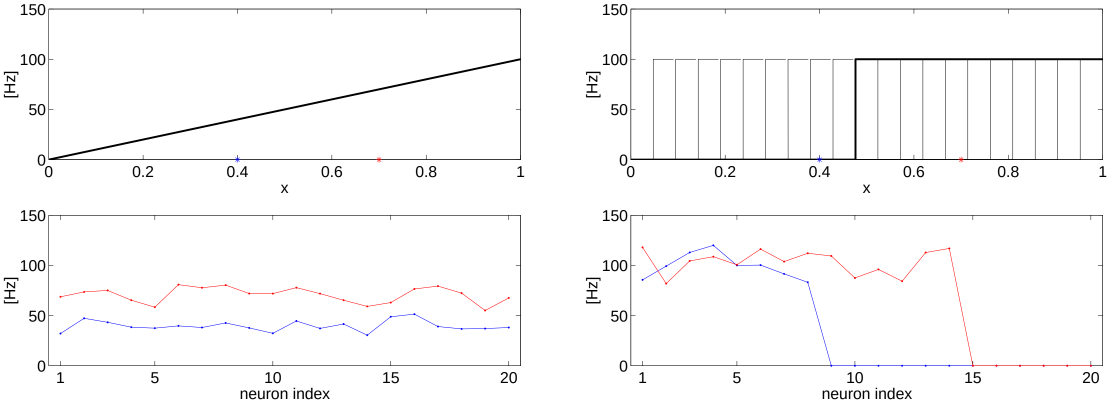

Figure 6.1. Intensity vs label-pattern coding. In case of intensity coding, neurons have rather linear tuning functions (left, upper), and averaging over the population leads to a reliable representation of the signal (left, lower). In the opposite extreme of label-pattern coding, the tuning functions are binary (right, upper) and the signal is encoded by the binary, spatial pattern of neuronal activity (right, lower). For illustration, the activations caused by two different stimuli (blue and red) are shown.

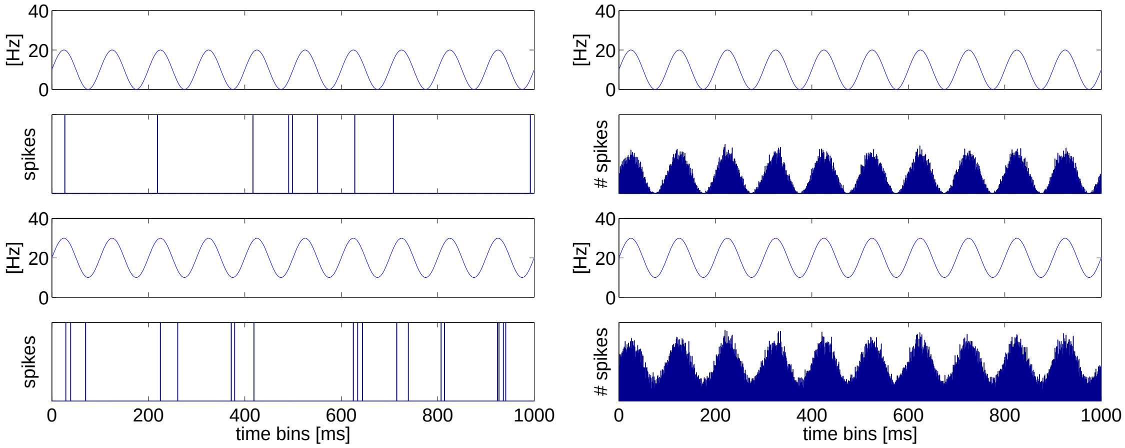

Figure 6.2. Intensity estimation from Poisson spike trains via pooling. While in case of a single neuron (left) an intensity estimate is very unreliable, averaging over the population allows for analog signaling with high temporal precision (right).

and the objective function reads

$$\begin{aligned}\chi^2[\mu_1(x),\ldots,\mu_N(x)] &= E[x^2]-E[\hat{x}^2] \qquad (6.5)\\ &= \frac{1}{3}-\sum_{k_1=0}^{\infty}\cdots\sum_{k_N=0}^{\infty}\frac{\left(\int_0^1 x\,p(\mathbf{k}|\mu_1(x),\ldots,\mu_N(x))\,dx\right)^2}{\int_0^1 p(\mathbf{k}|\mu_1(x),\ldots,\mu_N(x))\,dx}\,.\end{aligned}$$

Since the multi-dimensional integration in Eq. 6.5 cannot be solved analytically, Monte-Carlo methods have been used to evaluate χ^2 (see appendix A.2). In this way, χ^2 is determined as a function of μ_{max} for four different ramp coding schemes ($\lambda = 0, \frac{1}{N}, \frac{1}{2}, 1$) and three different wave coding schemes $(1,2,3,4,5,6,7,8,9,10)$, $(1,1,2,2,3,3,4,4,5,5)$, and $(1,1,2,2,3,3,4,4,5,6)$ in case of $N=10$ (see Fig. 6.4).

The ramp coding schemes with $\lambda > \frac{1}{N}$ are always worse than the ramp coding scheme with $\lambda = \frac{1}{N}$, which is optimal with respect to Fisher information within this class (see Chp. 5). The ramp coding scheme with $\lambda = 0$, however, is slightly better for $\mu_{max} \lessapprox 5$, while it becomes worse for $\mu_{max} > 5$, which is again due to the bias that is unavoidable for discrete encodings. By the use of wave functions, however, this bias can be reduced exponentially fast with increasing number of neurons. The bias of the wave coding scheme $(1,2,3,4,5,6,7,8,9,10)$ e.g. is equal to $\frac{1}{12}2^{-20}$ so that this code has a smaller error than all codes of the class $\mathcal{R}_2^{10}$ within the region $6 \lessapprox \mu_{max} < \frac{3}{25}2^{18} \approx 31457$. But also for $\mu_{max} < 6$ it is possible to achieve a much smaller error with binary encoding, if one introduces some redundancy in the wave coding scheme. A simple choice e.g. is the code $(1,1,2,2,3,3,4,4,5,5)$, where each wave function exists twice. Of course this redundancy can only be used at the cost of a much larger bias, which now is $\frac{1}{12}2^{-10}$. The code $(1,1,2,2,3,3,4,4,5,6)$

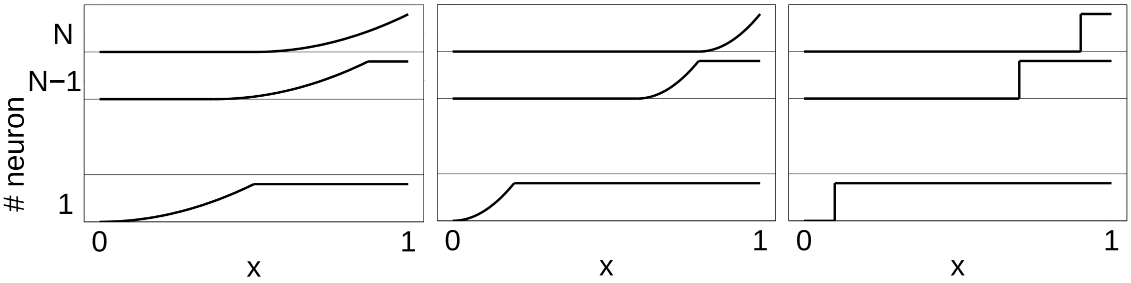

Figure 6.3. Sketch of the population ramp coding scheme $\mathcal{R}_2^N$ for $\lambda = \frac{1}{2}$ (left), $\lambda = \frac{1}{N}$ (middle), and $\lambda = 0$ (right). In case of $\lambda = 1$ the tuning functions of all neurons become identical with f^{asymp} (see Fig. 5.4 left).

is an example that illustrates, how one may achieve better compromises between redundancy and bias reduction with particular combinations of wave functions.

Taken together, the comparison of the MMSE for these examples confirms the hypothesis that the contribution of intensity coding to the coding accuracy of population codes is rather small. Another way to express this fact is that it appears that a certain required precision χ^2 with respect to a given μ_{max} can always be achieved with a minimal number of neurons, if the encoding is binary.

6.1 Encoding of a circular random variable

For better illustration, we also demonstrate the superiority of binary population coding for the frequently studied case of encoding a circular random variable $\phi \in [0, 2\pi)$ [TM91; SS93; SA94; ES97]. While it has been argued for the optimality of cosine tuning functions [SA94; Tod02], the results in section 5.5 suggest that box-shaped tuning functions are advantageous. In fact, it turns out that box tuning functions lead to a much smaller minimum mean squared error within a large range of μ_{max}.

For sufficiently large N the MMSE of cosine tuning functions

$$f_j^{\cos}(\phi) = \frac{f_{\max} + f_{\min}}{2} + \frac{f_{\max} - f_{\min}}{2} \cos(\phi - c_j) \quad , j = 1, \ldots N \tag{6.6}$$

with c_j uniformly distributed over $[0, 2\pi)$ is approximately equal to the inverse of the total average Fisher information [SS93]

$$J = \frac{N}{2\pi} \int_0^{2\pi} T \frac{(f_j^{\cos\prime}(\phi))^2}{f_j^{\cos}(\phi)} d\phi \leq \frac{\mu_{max}}{2} N \, , \tag{6.7}$$

where equality holds at the inequality, if there is no background noise (i.e. $f_{\min} = 0$). Hence, the resulting MMSE of cosine tuning functions is not smaller than $2/(N\mu_{max})$.

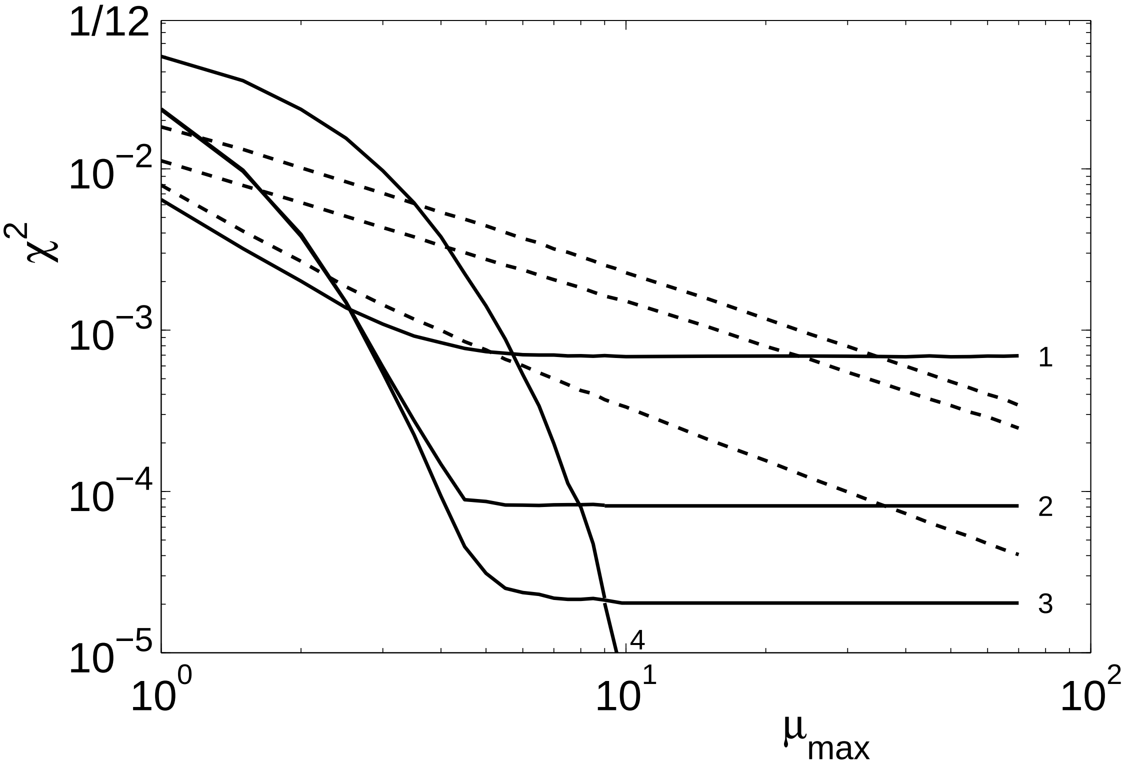

Figure 6.4. Comparison of the MMSEs of different population coding schemes in case of $N = 10$ neurons (log-log-scale). Solid lines are associated with pure binary encodings, while dashed lines indicate the MMSE of R_2^{10} in case of $\lambda = 1$ (upper), $\lambda = \frac{1}{2}$ (middle), $\lambda = \frac{1}{N}$ (lower). The range of $\frac{1}{N} \leq \lambda \leq 1$ corresponds to those tuning function arrays in R_2^N that are one-to-one. Within this subclass the choice of the minimal λ is optimal for all μ_{max}. The different binary encodings are distinguished by a number (1-4) at the end of the graphs. The MMSE of R_2^{10} in case of $\lambda = 0$ (solid, 1) is slightly better than the case of $\lambda = \frac{1}{N}$ until it saturates at about $\mu_{max} \approx 5$. The square wave function encodings (1,1,2,2,3,3,4,4,5,5) (solid, 2) and (1,1,2,2,3,3,4,4,5,6) (solid, 3) that exhibit some amount of redundancy achieve a substantially smaller MMSE within $3 \lessapprox \mu_{max} \lessapprox 8$. The coding scheme (1,2,3,4,5,6,7,8,9,10) (solid, 4), which has no redundancy and a minimal maximum frequency, is by far the best encoding for $10 \lessapprox \mu_{max} \lessapprox 31457$.

On the other hand, when the tuning functions are box shaped (i.e. if $\cos(\phi - c_j)$ is replaced with $\mathrm{sgn}(\cos(\phi - c_j))$ in Eq. 6.6), no unbiased estimator exists and even in the absence of noise the error cannot be smaller than the minimal bias, which is given by

$$bias = \frac{1}{12}\left(\frac{\pi}{N}\right)^2 . \tag{6.8}$$

However, the error variance decreases substantially faster in case of binary tuning functions than in case of smooth tuning functions. This fact suggests how to derive a rule of thumb for the critical μ_{max}, namely by equating the minimum bias of the

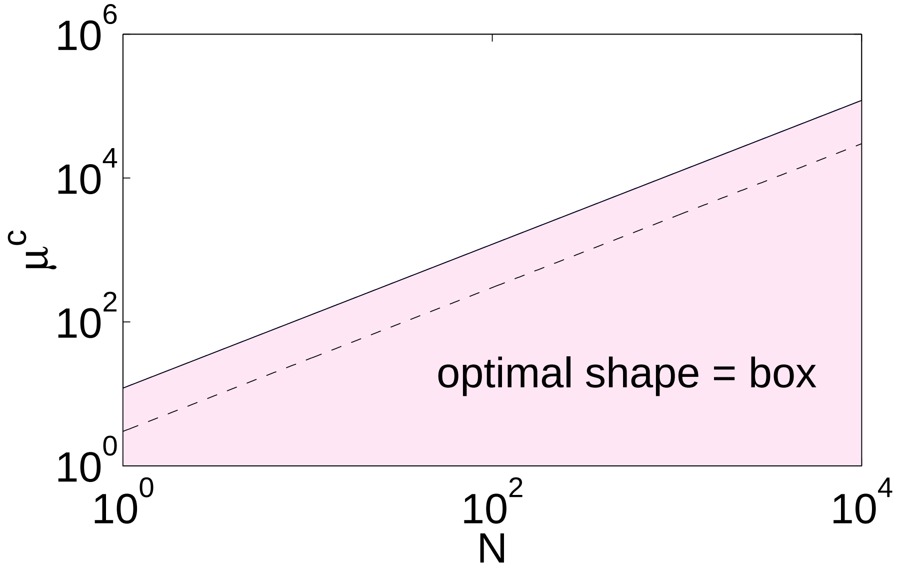

Figure 6.5. Phase diagram: box vs cosine tuning. For small χ^2 the MMSE in case of cosine tuning can be bounded from below by the inverse Fisher information, which yields $(\chi^2)^{cos} \geq 1/(N\bar{\mu})$. The intersection of this bound with the limiting value $\lim_{\bar{\mu}\to\infty} (\chi^2)^{box} = 1/(12N^2)$ provides a lower bound for the critical $\bar{\mu}$, up to which the box code is optimal (see figure, solid). A similar lower bound holds for all smooth tuning functions (dashed).

box encoding to the lower bound $1/J$ on the MMSE of cosine tuning functions, which yields:

$$\mu^c_{max} = \frac{24}{\pi^2} N \,. \tag{6.9}$$

In order to support the validity of Eq. 6.9, we determined the reconstruction error for the box coding scheme numerically in case of $\mu^c_{max}/2$ (i.e. higher noise level) with a rather large background noise $\mu_{min} = \mu^c_{max}/20$. While the lower bound (6.7) for the cosine coding scheme is twice as large at $\mu^c_{max}/2$ than at μ^c_{max}, the mean squared error of a simple perceptron estimator has already reached the lower bound (6.8) for the box tuning, provided N is sufficiently large (i.e. $N \gtrsim 20$). The main result is displayed by the phase diagram (Fig. 6.5), which indicates the region within which box tuning leads to a smaller coding error than cosine tuning. From this diagram, it is obvious that label-pattern coding is the more relevant, the larger the size of the population.

Note, that the lower bound (6.8) can only be reached for particular arrangements of the tuning function centers. If the centers are uniformly distributed over the ring, the average bias is smaller than $2\pi^2/N^2$, which gives reason to use $\mu^c_{max} = N/\pi^2$ as a rule of thumb instead. The scaling of this rule with respect to N remains the same as in Eq. 6.9 though.

Chapter 7

Linear Decoding

Until now, we have studied neuronal encodings from the perspective of an “ideal” observer. The MS-estimator, in particular, has to be considered as such in the sense of minimizing the average mean squared error. As mentioned in (3.5.2), there is, however, reason to assume constraints on the computational power of subsequent neuronal readout. Therefore, this chapter investigates how the coding accuracy is affected if the decoding is constrained to be linear. Linear estimation is interesting also because of its analytical tractability. Therefore, it can be used as an upper bound on the MMSE, which is not sharp, but of unrestricted validity at least. Since the neuronal readout is likely to be less constrained than in the case of being linear, the restriction to linear estimates can be seen as the limiting case of a maximally constrained decoding, as opposed to the case of the entirely unrestricted MS-estimator.

7.1 LMMSE and label-pattern coding

In general, the *linear minimum mean square estimator* (*LMS-estimator*) of a random variable x [Kay93] is given by

$$\hat{x}(\vec{y}) := \mathrm{E}\left[x\right] + \sum_{j=1}^{N} a_j y_j = C_{xy} C_{yy}^{-1} (\vec{y} - \mathrm{E}\left[\vec{y}\right]) \quad . \tag{7.1}$$

There, $(C_{xy})_i = E[xy_i] - E[x]E[y_i]$ is the covariance between x and the observable random variables $y_1, \ldots, y_N$ and

$$(C_{yy})_{ij} = E[y_i y_j] - E[y_i]E[y_j] = \begin{cases} \mathrm{E}\left[\mathrm{E}\left[y_i y_j \middle| x\right]\right] & , \quad i \neq j \\ \mathrm{Var}\left[y_j\right] = \mathrm{Var}\left[\mathrm{E}\left[y_i \middle| x\right]\right] + \mathrm{E}\left[\mathrm{Var}\left[y_i \middle| x\right]\right] & , \quad i = j \end{cases}$$

is the covariance between y_i and y_j. The optimality of this estimator can simply be shown by using the Hilbert space structure of zero mean random variables, for which the scalar product is defined by the covariance. Accordingly, the mean squared error

$$E[(x - \hat{x}(\vec{y}))^2] = \left|\left| x - \mathrm{E}\left[x\right] + \sum_{j=1}^{N} a_j (y_j - \mathrm{E}\left[y_j\right]) \right|\right|^2 \tag{7.2}$$

corresponds to the squared length of the error vector $(x - \hat{x}(\vec{y}))$, which takes its minimum, if

$$E[(x - \hat{x}(\vec{y}))(y_j - \mathrm{E}\left[y_j\right])] = 0 \tag{7.3}$$

for all $j = 1, \ldots, N$, because then the error vector cannot be reduced any further. Since Eq. 7.3 can also be written as a matrix equation

$$C_{xy} - C_{yy}\vec{a} = \vec{0} \quad , \tag{7.4}$$

the coefficients $\vec{a}$ of the optimal linear estimator are obtained simply by rearranging this equation.

Now we want to use the optimal linear estimator to reconstruct a signal from a population code of N neurons, for which the sensitivity profiles are described by $\mu_j(x) := E[y_j|x]$, $j = 1, \ldots, N$. In case the noise of different neurons is mutually uncorrelated (i.e. $E[y_i y_j|x] = E[y_i|x]E[y_j|x]$), the components of the covariance matrix C_{yy} always have the following form:

$$(C_{yy})_{ij} = E[\mu_i(x)\mu_j(x)] - E[\mu_i(x)]E[\mu_j(x)] + \delta_{ij}\, \bar{\nu}_j^2 \tag{7.5}$$

where $\bar{\nu}_j^2 := E[Var[y_j|x]]$ denotes the mean of the conditional variance of y_j (i.e. the noise level). Hence, the noise model only affects the main diagonal of C_{yy} and nothing but the average variance of the noise distribution matters. To give some relevant examples, for constant additive noise (i.e. $Var[y_j|x]$ does not depend on x) it holds

$$\bar{\nu}_j^2 = Var[y_j|x]\,. \tag{7.6}$$

In case of a Poisson distribution, we obtain

$$\bar{\nu}_j^2 = E[\mu_j(x)] \tag{7.7}$$

and for a Bernoulli distribution it holds

$$\bar{\nu}_j^2 = E[\mu_j(x)(1 - \mu_j(x))]\,. \tag{7.8}$$

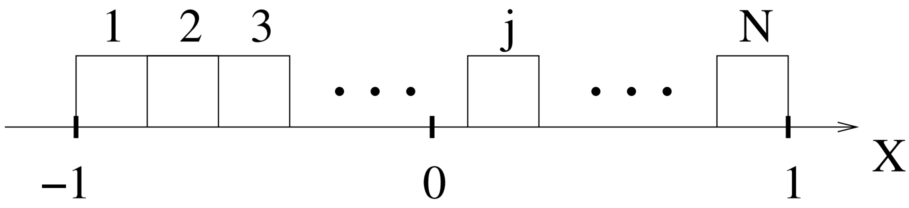

Figure 7.1. Sketch of box tuning function array. The intervals of positive firing rates of the different tuning functions constitute an equipartition of the support of X.

Now, let us assume that x is uniformly distributed $x \sim U(-1,1)$ and the generalized[1] tuning functions are given by

$$\begin{aligned} \mu_j(x) &= \bar{\mu}\Theta(\vartheta_j, x, \vartheta_j + w) && (7.9)\\ w &= \frac{2l}{N}, \, l \in \{1,\dots,N\} && (7.10)\\ \vartheta_j &= (2-w)\frac{j-1}{N-1} - 1 && (7.11) \end{aligned}$$

Due to the similarity of the tuning functions, the average noise $\bar{\nu}_j^2 = \bar{\nu}^2$ becomes independent of j, and the corresponding covariance matrix reads

$$(C_{yy})_{ij} = \bar{\mu}^2 \left\{ [l - |i-j|]_+ \frac{l}{N} - \left(\frac{l}{N}\right)^2 \right\} + \delta_{ij}\bar{\nu}^2 \, . \qquad (7.12)$$

For the sake of clarity, we now restrict the analysis to the case of $l = 1$, for which the array of tuning functions is sketched in Fig. 7.1. In this case, the covariance matrix can be written as

$$(C_{yy})_{ij} = \delta_{ij}\frac{\bar{\mu}^2 + N\bar{\nu}^2}{N} - \left(\frac{\bar{\mu}}{N}\right)^2 \, . \qquad (7.13)$$

Since generally for any $N \times N$ matrix of the form $(\delta_{ij}a + b)$ the inverse is given by $(\delta_{ij}A + B)$ with

$$A = \frac{1}{a} \text{ and } B = \frac{b}{Nab + a^2} = \frac{1}{Na + a^2/b} \qquad (7.14)$$

[1] The tuning functions $\mu_j(x)$ do not have the dimension of a firing rate, but directly represent the mean spike count $\mu_j(x) = Tf_j(x)$, which is the quantity that effectively matters. Therefore, I call the $\mu_j(x)$ 'generalized tuning functions', since the term 'tuning function' is commonly used for the mean firing rates.

C_{yy}^{-1} has a short representation for this example

$$(C_{yy}^{-}1)_{ij} = \delta_{ij}\frac{N}{\bar{\mu}^2 + N\bar{\nu}^2} + B \quad . \tag{7.15}$$

For uncorrelated noise, the general form of the covariance matrix C_{xy} is given by

$$(C_{xy})_j = E[x\mu_j(x)] - E[x]E[\mu_j(x)] \tag{7.16}$$

In case of the example (Eq. 7.9) this yields

$$(C_{xy})_j = \frac{\bar{\mu}}{2}\left(1 - \frac{w}{2}\right)\left(2\frac{j-1}{N-1} - 1\right)w \overset{(l=1)}{=} \frac{\bar{\mu}}{N}\frac{N-1}{N}\left(2\frac{j-1}{N-1} - 1\right) \tag{7.17}$$

Hence, the coefficients of the optimal estimator are given by

$$a_j = (C_{xy}C_{yy}^{-1})_j = \frac{NC_{xy}}{\bar{\mu}^2 + N\bar{\nu}^2} + \underbrace{B\sum_{i=1}^{N}(C_{xy})_i}_{=0} = \frac{\bar{\mu}}{\bar{\mu}^2 + N\bar{\nu}^2}\frac{N-1}{N}\left(2\frac{j-1}{N-1} - 1\right) \tag{7.18}$$

Substituting $\hat{x} = C_{xy}C_{yy}^{-1}$ in Eq. 7.2 leads to the general expression for the *linear minimum mean squared error* (*LMMSE*):

$$\chi^2_{LMS} \equiv E[(x - \hat{x}(\vec{y}))^2] = Var[x] - C_{xy}^T C_{yy}^{-1} C_{xy} = Var[x] - \hat{x}C_{xy} \tag{7.19}$$

Substituting Eq. 7.15 and Eq. 7.17 yields (see appendix 7.2)

$$\chi^2_{LMS} = \frac{1}{3} - \frac{\bar{\mu}}{\bar{\mu}^2 + N\bar{\nu}^2}\frac{\bar{\mu}}{N}\left(\frac{N-1}{N}\right)^2 \sum_{j=1}^{N}\left(2\frac{j-1}{N-1} - 1\right)^2 \tag{7.20}$$

$$= \frac{1}{3}\left\{\frac{N\nu^2}{N\nu^2 + \bar{\mu}^2} + \frac{1}{N^2}\cdot\frac{\bar{\mu}^2}{\bar{\mu}^2 + \nu^2 N}\right\} . \tag{7.21}$$

Note that $\bar{\nu}^2 = \bar{\nu}^2(N)$ depends in general on the encoding and hence it may depend on N. If $\mathrm{Var}\,[y_i|\, x]$ is independent of x (i.e. additive noise), it holds $\bar{\nu}^2 = \mathit{const}$. In that case, the second term at the r.h. side of (7.21) scales as N^{-2}, while the first term within the brackets *increases* with N and converges to 1. In other words, for additive noise the quadratic information gain of the LMS-estimator is zero in the limit $N \to \infty$. This is different in case of Poisson noise, for which $\bar{\nu}^2 = \frac{\bar{\mu}}{N}$ so that Eq. 7.21 becomes

$$\chi^2_{LMS} = \frac{1}{3}\left\{\frac{1}{1+\bar{\mu}} + \frac{1}{N^2}\cdot\frac{\bar{\mu}}{1+\bar{\mu}}\right\} = \frac{1}{3}\cdot\frac{1}{1+\bar{\mu}}\left(1+\frac{\bar{\mu}}{N^2}\right) \geq \frac{1}{3}\cdot\frac{1}{1+\bar{\mu}}\,. \tag{7.22}$$

This means that for the Poisson noise model the average risk is a strictly decreasing function of N. However, also in this case, the mean squared error remains finite for all N, and the lower bound $1/(1+\bar{\mu})$ depends only on the maximum firing rate $\bar{\mu}$.

There is, however, a way to increase $\bar{\mu}$ effectively by using a number of identical copies, say R, of the tuning function array and to sum over the responses of all neurons that have identical tuning curves. Each pool of equivalent neurons then has an effective maximum mean spike count $\bar{\mu}^{pool} = R\bar{\mu}$, but the total number of neurons required to obtain N^{pool} sub-populations clearly increases as well (i.e. $N = N^{pool}R$). For such a (more) redundant encoding strategy Eq. 7.22 reads

$$\begin{aligned}
\chi^2_{LMS} &= \frac{1}{3}\cdot\frac{1}{1+\bar{\mu}^{pool}}\left(1+\frac{\bar{\mu}^{pool}}{(N^{pool})^2}\right)\\
&= \frac{1}{3}\cdot\frac{1}{1+\bar{\mu}R}\left(1+\frac{\bar{\mu}R^3}{N^2}\right)\\
&= \frac{1}{3}\cdot\frac{1}{1+\frac{1}{\bar{\mu}R}}\left(\frac{1}{\bar{\mu}R}+\frac{R^2}{N^2}\right) \leq \frac{1}{3}\left(\frac{1}{\bar{\mu}R}+\frac{R^2}{N^2}\right)\,.
\end{aligned} \tag{7.23}$$

The upper bound suggests to choose an optimal redundancy level R^{opt} according to

$$\frac{1}{\bar{\mu}R_{opt}} = \frac{R^2_{opt}}{N^2} \quad . \tag{7.24}$$

Consequently, the average risk for the choice $R_{opt} = \sqrt[3]{\frac{N^2}{\bar{\mu}}}$ yields

$$\chi^2_{LMS} \leq \frac{2}{3}(N\bar{\mu})^{-\frac{2}{3}} \tag{7.25}$$

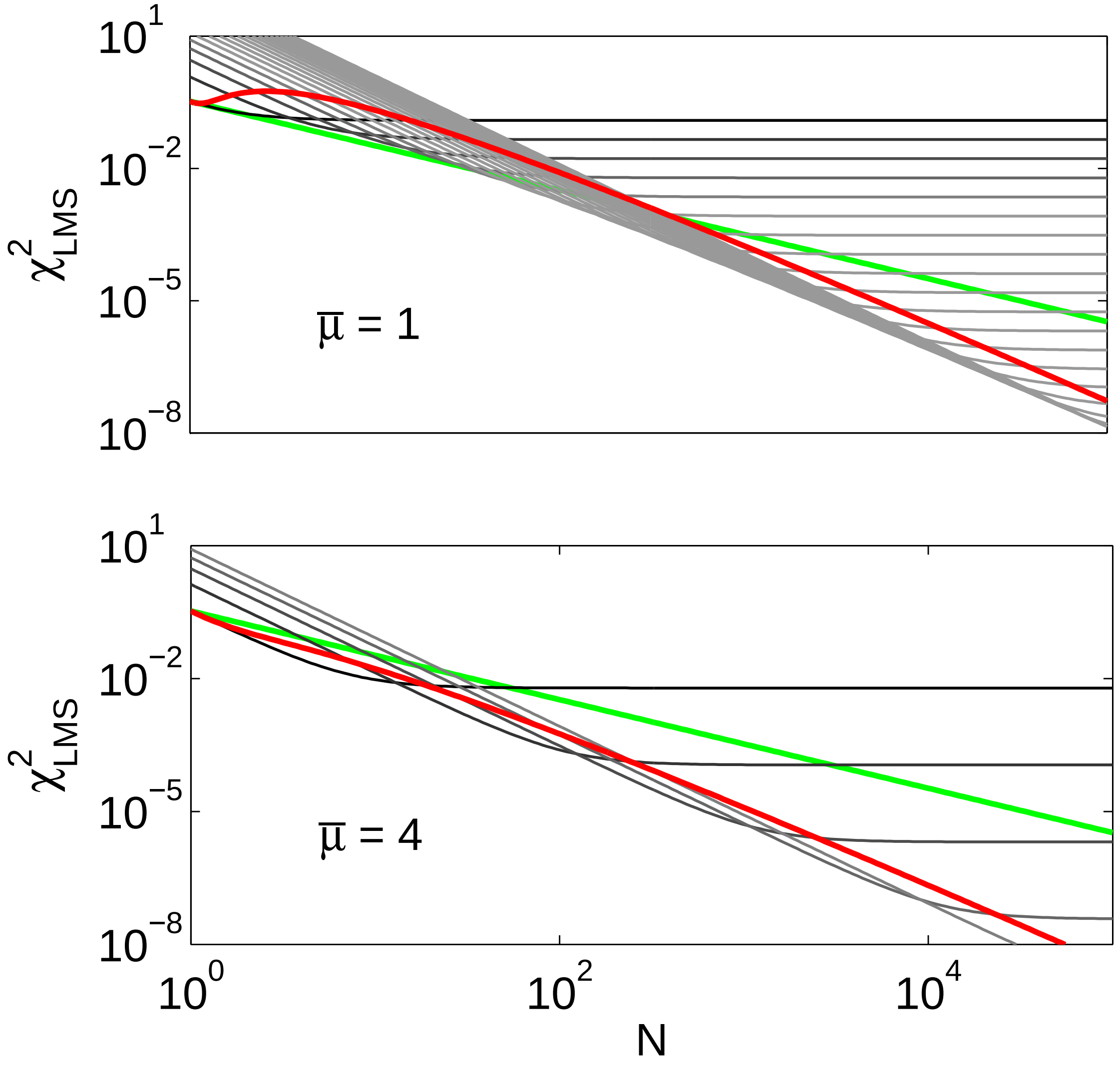

Figure 7.2. The LMMSE as a function of N depends on the chosen redundancy level. For illustration these functions are shown for $R = 1, 2, 3, \ldots$ (grey, from black to pale) in case of $\bar{\mu} = 1$ (upper) and $\bar{\mu} = 4$ (lower). The choice $R(N) = \frac{2}{\bar{\mu}} \log N$ is close and 'parallel' to the lower envelope of these functions.

This means that even if one selects an optimal level of redundancy, the LMMSE of the box encoding strategy decreases more slowly than $\frac{1}{N}$. Since the latter is clearly achievable in case of linear tuning functions, one might hypothesize that computational constraints on the subsequent neuronal readout support the idea of intensity coding. In fact, inverse linear scaling is commonly the best that one may achieve with linear decoding and a finite amount of noise in the limit $N \to \infty$.

However, as mentioned before, the case of linear decoding is likely to be overconstrained. Therefore, it is instructive to see that even simplest nonlinearities are sufficient to achieve a scaling close to $\frac{1}{N^2}$, which is the optimum in case of the unconstrained MS-estimator. To this end, the Bernoulli noise model is of interest as it can be used in order to mimic a simple saturation non-linearity.

Now, the response of each pool corresponds to a saturating sum, which is equal to one, if at least one neuron emits a spike, and zero otherwise. This picture is not restricted to Bernoulli neurons, but can be applied to any neuron model. It is only required to specify the probability $p_{\geq 1}$ of the spike count to be larger than or equal to one. Then the maximum pool response probability is given by

$$\bar{\mu}^{pool} = 1 - (1 - p_{\geq 1})^R , \tag{7.26}$$

which scales exponentially with R. In case of Poisson neurons it holds $p_{\geq 1} = 1 - \exp(-\bar{\mu})$, leading to the simple expression

$$\bar{\mu}^{pool} = 1 - \exp(-R\bar{\mu}) \quad . \tag{7.27}$$

Substituting $\bar{\nu}^2 = (1 - \bar{\mu})\bar{\mu}/N$ for the Bernoulli noise model into Eq. 7.21 yields

$$\chi^2_{LMS} = \frac{1}{3}\left(1 - \bar{\mu}^{pool} + \frac{\mu^{pool}}{(N^{pool})^2}\right) . \tag{7.28}$$

Inserting (7.27) into (7.28) and using $N^{pool} = N/R$ we finally obtain for the LMMSE

$$\chi^2_{LMS} = \frac{1}{3}\left(\exp(-R\bar{\mu}) + (1 - \exp(-R\bar{\mu}))\frac{R^2}{N^2}\right) . \tag{7.29}$$

Using a redundancy level $R = \frac{2}{\bar{\mu}} \log N$, the error scales proportional to $\left(\frac{\log(N)}{N}\right)^2$, which is close to $\frac{1}{N^2}$ and substantially faster than $\frac{1}{N}$. In Fig. 7.2 it is shown that the particular choice $R = \frac{2}{\bar{\mu}} \log N$ is close to being optimal.

In conclusion, this example demonstrates that even very simple non-linearities allow for a substantial improvement of signal transmission and suggest to consider other encoding strategies than is the case for strictly linear decoding.

7.2 Appendix

For the derivation of (7.20) from (7.21) it is used that

$$
\begin{aligned}
& \sum_{j=1}^{N} \left(2\,\frac{j-1}{N-1} - 1 \right)^2 \\
= \; & \sum_{j=1}^{N} \left\{ 4 \left(\frac{j-1}{N-1} \right)^2 - 4\,\frac{j-1}{N-1} + 1 \right\} \\
= \; & \sum_{j=1}^{N} \left\{ 4\,\frac{j^2 - 2j + 1}{(N-1)^2} - 4\,\frac{j-1}{N-1} + 1 \right\} \\
= \; & \sum_{j=1}^{N} \left\{ \frac{4}{(N-1)^2}\, j^2 - \left(\frac{8}{(N-1)^2} + \frac{4}{N-1} \right) j \right\} + N \left(\frac{4}{(N-1)^2} + \frac{4}{N-1} + 1 \right) \\
= \; & \frac{4}{(N-1)^2} \left(\sum_{j=1}^{N} j^2 \right) - \frac{8 + 4(N-1)}{(N-1)^2} \left(\sum_{j=1}^{N} j \right) + N\, \frac{(N-1)^2 + 4(N-1) + 4}{(N-1)^2} \\
= \; & \frac{4}{(N-1)^2} \cdot \frac{N(N+1)(2N+1)}{6} - \frac{4N+4}{(N-1)^2} \cdot \frac{N(N+1)}{2} + \frac{N((N-1)+2)}{(N-1)^2} \\
= \; & \frac{4}{6} \cdot \frac{N(N+1)^2 + N^2(N+1)}{(N-1)^2} - 2\,\frac{N(N+1)^2}{(N-1)^2} + \frac{N(N+1)^2}{(N-1)^2} \\
= \; & \frac{4}{6} \cdot \frac{N^2(N+1)}{(N-1)^2} + \frac{4-6}{6} \cdot \frac{N(N+1)^2}{(N-1)^2} \\
= \; & \frac{2N^3 + 2N^2 - N3 - 2N^2 - N}{3(N-1)^2} = \frac{N(N+1)(N-1)}{3(N-1)^2} = \underline{\underline{\frac{N(N+1)}{3(N-1)}}} \,.
\end{aligned}
$$

Substituting this into Eq. 7.20 yields

$$\begin{aligned}
\frac{1}{3}-\frac{\bar{\mu}}{\bar{\mu}^2+N\bar{\nu}^2}\frac{\bar{\mu}}{N}\left(\frac{N-1}{N}\right)^2\frac{N(N+1)}{3(N-1)} &= \frac{1}{3}\left\{1-\frac{\bar{\mu}}{\bar{\mu}^2+N\bar{\nu}^2}\left(\frac{N-1}{N}\right)^2\frac{N+1}{N-1}\right\} \\
&= \frac{1}{3}\left\{1-\frac{\bar{\mu}}{\bar{\mu}^2+N\bar{\nu}^2}\frac{(N-1)(N+1)}{N^2}\right\} = \frac{1}{3}\left\{1-\frac{\bar{\mu}}{\bar{\mu}^2+N\bar{\nu}^2}\left(1-\frac{1}{N^2}\right)\right\} \\
&= \frac{1}{3}\left\{\frac{N\bar{\nu}^2}{\bar{\mu}^2+N\bar{\nu}^2}+\frac{\bar{\mu}^2}{\bar{\mu}^2+N\bar{\nu}^2}\cdot\frac{1}{N^2}\right\} = \frac{1}{3N^2}\cdot\frac{\bar{\mu}^2+\nu^2N^3}{\bar{\mu}^2+\nu^2N} \\
&= \underline{\underline{\frac{1}{3}\left\{\frac{1}{1+\frac{\bar{\mu}^2}{N\nu^2}}+\frac{1}{N^2}\cdot\frac{\bar{\mu}^2}{\bar{\mu}^2+\nu^2N}\right\}}}.
\end{aligned}$$

Chapter 8

Optimality of Binary Coding: A Phase Transition

The previous four chapters investigated population coding from a very abstract point of view, and some readers may feel a little bit lost because of the variety of aspects that have been addressed. Therefore, I think, it is a good time to shortly recap the major lessons we have learned so far: Starting from the context of the literature on population coding, we saw in chapter 4 that the Fisher information of population codes is influenced by the tuning width and the dynamic range independently of each other.

While this result allowed to resolve the conflicting conclusions in the literature about the optimal tuning width, we noticed several difficulties at interpreting the underlying model in a meaningful way. Irrespective of the technical problems due to the limited validity of Fisher information, the preselection of the candidate encodings (chapter 5 and 6) as well as the choice of the decoder (chapter 7) have been shown to influence the shape of the optimal encodings in a crucial way. Three insights, in particular, shall be pointed out here:

- Strictly linear decoding is too restrictive, because simple saturation or threshold nonlinearities are sufficient to take advantage of label-pattern coding. This, in particular, makes a big difference for the shape of the optimal encoding.
- It is not possible to justify a preselection of tuning functions independent of a particular problem at hand. Therefore, it is important to note that the frequent restriction to bell-shaped tuning profiles and radial symmetric tuning functions impair the coding accuracy substantially.
- Minimization of the dynamic range turned out to be crucial for maximizing Fisher information in general and has been proven to be advantageous in several case studies with respect to the MMSE as well.

The purpose of the previous chapters was mainly to facilitate intuitions as to how the shape of an efficient neuronal representation is affected by a variety of mutually influencing factors. Here, this knowledge shall be used to derive a constraint on neural rate coding that directly applies to efficient coding models of natural images.

8.1 Multi-dimensional encoding in the visual system

The original work on the optimal tuning width in population coding by Hinton [Hin81] was motivated by a search for theoretical determinants of the coding efficiency of sensory neurons in the visual system. Accordingly, his study is related to the task of reconstructing the location of a single light dot in the two dimensional visual field from the neuronal responses. While Hinton already remarked that his scaling rule holds true only under the assumption that not more than one light dot is located within the same receptive field simultaneously, the more general setting has hardly been investigated so far.

In order to describe arbitrary visual stimuli, we use a scalar intensity field $I : [-b_x, b_x] \times [-b_y, b_y] \rightarrow [0, 1], (x, y) \mapsto I(x, y)$. Furthermore, we model the rate response of neuron $j = 1, \ldots, N$ by a memoryless linear-nonlinear cascade neuron model [SPPS04]

$$\begin{aligned} z_j[I] &= \langle F_j, I \rangle = \int_{-b_x}^{b_x} \int_{-b_y}^{b_y} F_j(x, y) I(x, y) dy dx && (8.1) \\ f_j[I] &= g_j(z_j[I]) && (8.2) \end{aligned}$$

where the first equation is a scalar product and $g_j(z_j)$ is some nonlinear gain function with limited output range (i.e. $f_{min} \leq g_j(z) \leq f_{max}$). In the particular case of Poisson noise, it is also referred to this model as the *linear-nonlinear-Poisson (LNP) model* [Chi01].

In case of $I(x, y)$ being a delta function (i.e. a 'dot stimulus'), this model becomes equivalent to those analyzed in previous studies discussed above, because then $f_j[I(x, y)] = F(x, y)$ becomes a tuning function for the location of the delta peak. Note, however, that radial symmetry or smoothness of F, is typically not conserved under the mapping Eq. 8.1, if $P(I)$ contains images other than delta functions. Even in the simple case, where I may be the sum of two delta functions, the stimulus space and the tuning functions become very different, because now a four-dimensional signal (x_1, y_1, x_2, y_2) has to be encoded. Clearly, the visual system actually deals with very high-dimensional data.

The merit of the linear-nonlinear cascade neuron model is that it allows to separate the ecological problem of optimizing the receptive fields with respect to the natural image statistics from the other, endogenous, problem to overcome the neuronal noise of rate signaling between neurons. In fact, this is a crucial step, because the optimal set of basis functions $\{F_j : j = 1, \ldots, N\}$ is mainly determined by the shape of the prior distribution $P(I)$, while the optimal gain function can be, to a large extent, considered independently. Furthermore, the degree of nonlinearity appears to be a plausible choice.

In summary, we here take the point of view that it makes sense to ask for the precision with which the N-dimensional vector $(z_1, \ldots, z_N)$ can be reconstructed from variable neuronal responses with mean $(g_1(z_1), \ldots, g_N(z_N))$, rather than to ask for the precision with which an intensity field I can be reconstructed under the assumption of a particular prior distribution $P(I)$. A different way to motivate this point of view is to make use of the hypothesis that each neuronal response may reflect an independent component of natural scenes [BS97; Ati92].

Under the assumption that the z_j are mutually independent (i.e. $P(z_1, \ldots, z_N) = \prod P(z_j)$), and that all dimensions are equivalent with respect to the variance of their prior distributions as well as to their relevance for further processing, it is optimal to encode each dimension exclusively by one neuron [BRP02]. Therefore, the study of optimal rate encoding of an individual neuron appears to be well suited to analyze the effect of neuronal noise on efficient neuronal representations.

This may come somewhat as a surprise, since it is widely believed that population coding is most relevant to understand the neural code. Although, in principle, it cannot be wrong to study populations instead of individual neurons, the usually studied case of point estimation with respect to the mean squared error loss in an Euclidian space is likely to misrepresent the actual neuronal functionality under natural conditions. For the special task of determining the position of a single light dot, it might be reasonable to consider a two-dimensional point estimation problem under squared error loss. However, this model becomes more or less irrelevant, if, in fact, the discrimination of natural scenes is important.

In conclusion, the optimization of population codes makes sense if the loss function implements behavioral relevance[1]. The loss functions used in contemporary models of optimal population coding, however, are not informed by behavioral relevance, but have been chosen merely for the sake of convenience. While strong knowledge is required to set up a meaningful loss function, the less ambitious question of how to signal a certain magnitude from one neuron to another, may be of relevance, independent of the particular task the neurons have to solve.

[1] This is the reason why the efficient coding principle puts so much emphasis on the role of natural stimuli.

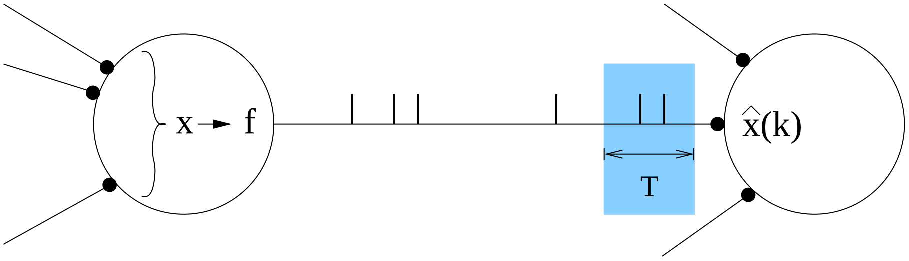

Figure 8.1. The presynaptic neuron at the left computes some analog value z from its synaptic inputs. In order to signal this quantity to another postsynaptic neuron, spikes are generated with a firing rate $g(z)$ and propagated to the postsynaptic neuron. The postsynaptic neuron integrates over all incoming spikes within a time window of length T. The resulting spike count k then serves as the basis for any computation of the postsynaptic neuron, for which an estimate $\hat{x}$ of z is required.

8.2 The rate coding bottleneck

The bottom line of the analysis in this chapter is to ask in how far the interpretation of rate coding as an analog code makes actually sense. In fact, the notion of rate coding is frequently linked to the idea of analog coding, which constitutes the basis for many neural network models. Apart from its hegemony in stimulus reconstruction experiments, the idea of rate coding can also be motivated by a basic biophysical property of neurons, namely the temporal integration of the postsynaptic cells over the current pulses induced by the presynaptic spikes. Reliable inference from the observed number of spikes about the underlying firing rate of a neuronal response, however, requires a sufficiently long time interval, while integration times of neurons *in vivo* [SK93] as well as reaction times of humans or animals when performing classification tasks [KXFP01; TFM96] are known to be rather short. Therefore, it is important to understand, how neural rate coding is affected by the limited time window, which in reality is available for decoding.

The motivating picture, we have in mind, refers to the communication process from one neuron to another, where we assume that the presynaptic neuron has computed an analog number z from its inputs and is now faced with the problem of signaling it over some distance along its axon to other neurons by the use of spikes (Fig. 8.1). In other words, z is assumed to represent exactly the "relevant information" encoded by the neuron, which need not match with the stimulus parameters commonly investigated experimentally. For this situation, we seek to determine optimal gain functions, such that the MMSE, with which z can be inferred by a postsynaptic neuron, is minimized. A gain function with respect to z constitutes a neuronal response function very similar to an f-I curve as known from experimental studies. The theoretical analysis in this chapter, however, does not rely on assumptions about particular physical signals corresponding to z. The essential question is simply, as to how it is possible to overcome the rate-coding bottleneck.

In 1996, Softky pointed out that there is a trade-off between the higher information content of each analog "message" and the lower rate at which this message may be sent [Sof96], so that the question arises how relevant the idea of analog coding actually is for neuronal processing in the brain. Although this is an important problem that even may be decidable experimentally, it did not receive much attention in neuroscience until today.

Here, we analyze this issue by seeking the optimal gain function that minimizes the MMSE for a uniform source signal transmitted through a Poisson channel as a function of the maximum mean number of spikes. In formal terms, the issue is to optimally encode a real random variable z in the number of pulses emitted by a neuron within a certain time window. Thereby, z stands for the analog signal computed by a presynaptic neuron that shall be transmitted to subsequent neurons. The neuronal output, actually read out by subsequent neurons, however, is given by the discrete number of spikes k integrated within a time interval of length T. The statistical dependency between z and k then is specified by the assumption of Poisson noise

$$p(k|\mu(z)) = \frac{(\mu(z))^k}{k!} \exp\{-\mu(z)\}\,, \tag{8.3}$$

and the choice of the gain function $g(z)$, which together with T determines the mean spike count $\mu(z) = Tg(z)$. An important additional constraint is the limited range of the neuronal firing rate, which can be included by the requirement of a bounded gain function ($g_{min} \leq g(z) \leq g_{max}\,, \forall\, x$). Since inhibition can reliably prevent a neuron from firing, we will consider the case $g_{min} = 0$ most of the time. Instead of specifying g_{max} it makes sense to impose a bound on the mean spike count directly (i.e. $\mu(z) \leq \mu_{max}$), because g_{max} constitutes a meaningful constraint only with respect to a fixed time window of length T. Since μ_{max} has a crucial effect on the signal-to-noise ratio, we will analyze the coding properties as a function of μ_{max}.

In reality g_{max} is bounded and fixed so that $\bar{\mu} = g_{max}T$ is directly related to the rate $1/T$ at which independent signals can be transmitted. Hence, in our study the trade-off between the higher information content of each analog "message" and the lower rate at which this message may be sent corresponds to the larger amount of time T that is necessary to achieve a lower distortion χ^2 by increasing the range of analog signaling. For the optimization of a gain function the MMSE χ^2 reads

$$\chi^2[\mu(z)] = \frac{1}{3} - \sum_{k=0}^{\infty} \frac{\left(\int_0^1 x\, p(k|\mu(z))\, dx\right)^2}{\int_0^1 p(k|\mu(z))\, dx}\,. \tag{8.4}$$

With respect to the issue of optimal analog rate signaling between neurons, the MMSE appears to be a well-suited objective function.

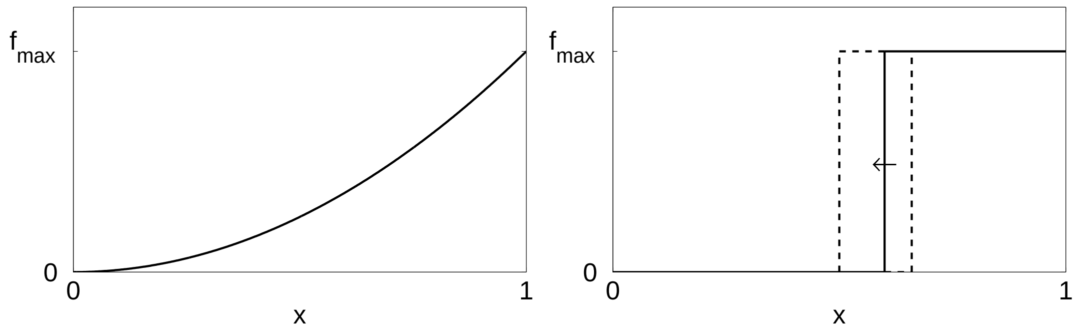

Figure 8.2. While the parabolic gain function g^{asymp} (left) is asymptotically optimal in the limit $\mu_{max} \to \infty$, the step function g^{binary} (right) is advantageous for small μ_{max}. The optimal threshold of the step function moves from $\frac{2}{3}$ to $\frac{1}{2}$ with increasing μ_{max}.

8.3 Optimal gain functions in the limiting cases

As derived above on the basis of Fisher information, the optimal gain in the asymptotic limit $T \to \infty$ has a parabolic shape (Fig. 8.2, left):

$$g^{asymp}(z) = ((\sqrt{g_{max}} - \sqrt{g_{min}})x + \sqrt{g_{min}})^2 \ . \tag{8.5}$$

For any finite μ_{max}, however, this gain function is not necessarily optimal. In particular, in the limit $\mu_{max} \to 0$, the Poisson distribution converges uniformly to a Bernoulli distribution with $P(k|\mu) = \mu^k(1-\mu)^{1-k}$ for $k \in \{0,1\}$, and for the latter, it is straightforward to show that the optimal gain function is a step function (Fig. 8.2, right).

$$g^{binary}(z) = g_{min} + (g_{max} - g_{min})\,\Theta\left(x - \vartheta_{g_{min}}(\mu_{max})\right) \ , \tag{8.6}$$

In case of $g_{min} = 0$ the optimal threshold $\vartheta_{g_{min}}(\mu_{max}) \in [1/2, 2/3]$ as a function of μ_{max} can be determined analytically

$$\vartheta_0(\mu_{max}) = 1 - \frac{3 - \sqrt{8e^{-\mu_{max}} + 1}}{4(1 - e^{-\mu_{max}})} \tag{8.7}$$

as well as the corresponding MMSE (see appendix 8.7):

$$\chi^2[g^{binary}] = \frac{1}{12}\left(1 - \frac{3\,\vartheta_0^2(\mu_{max})}{[(1-\vartheta_0(\mu_{max}))(1-e^{-\mu_{max}})]^{-1} - 1}\right) \ . \tag{8.8}$$

The MMSE of the asymptotically optimal gain function is given by

$$\chi^2[g^{asymp}] = \frac{1}{3} - \frac{1}{2\,(\sqrt{\mu_{max}})^3} \sum_{k=0}^{\infty} \frac{1}{k!} \frac{\Gamma^2_{0,\mu_{max}}(k+1)}{\Gamma_{0,\mu_{max}}(k+\frac{1}{2})} \quad , \tag{8.9}$$

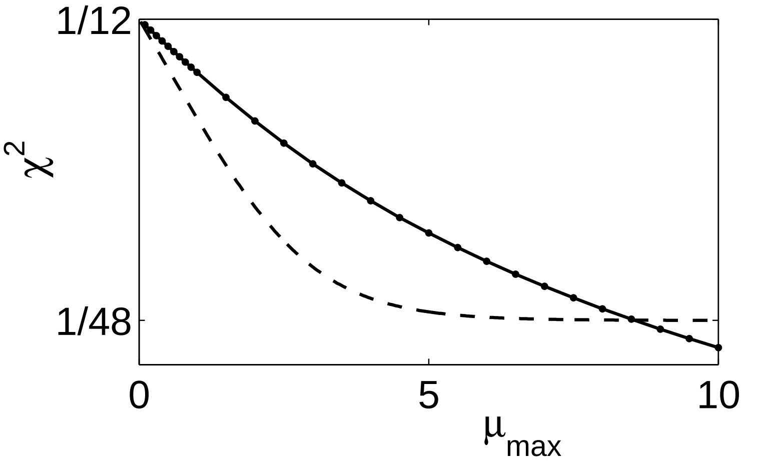

Figure 8.3. Comparison of the minimum mean squared error for the parabolic gain function (solid) and for the step function (dashed). The χ^2-axis has a logarithmic scale. In the relevant region $1 \leq \mu_{max} \leq 5$ the step function g^{binary} is clearly advantageous.

where $\Gamma_{r,s}$ denotes the truncated Gamma function

$$\Gamma_{r,s}(k) = \int_r^s t^{k-1} e^{-t} dt \quad . \tag{8.10}$$

A comparison of $\chi^2[g^{asymp}]$ with $\chi^2[g^{binary}]$ shows that the step function leads, in fact, to a smaller average reconstruction error than the parabolic gain function (see Fig. 8.3) if $\mu_{max} < 8.2$.

8.4 Numerically optimized gain functions for finite μ_{max}

The binary shape for small $\bar{\mu}$ and the continuous parabolic shape for large $\bar{\mu}$ implies that there has to be a transition from discrete to analog encoding when $\bar{\mu}$ is increased. Unfortunately, it is not possible to determine the optimal gain function within the entire set of all bounded functions $\mathcal{B} := \{f | f : [0, 1] \to [0, g_{max}]\}$, using the calculus of variations. Instead, we will have to chose certain parameterized function spaces in advance that are feasible for the optimization. In order to not rely on a particular parameterization, we investigated a variety of function classes for which the most important are presented in the following.

Let us first consider the classes $\mathcal{S}_\lambda$ of piecewise constant staircase functions with

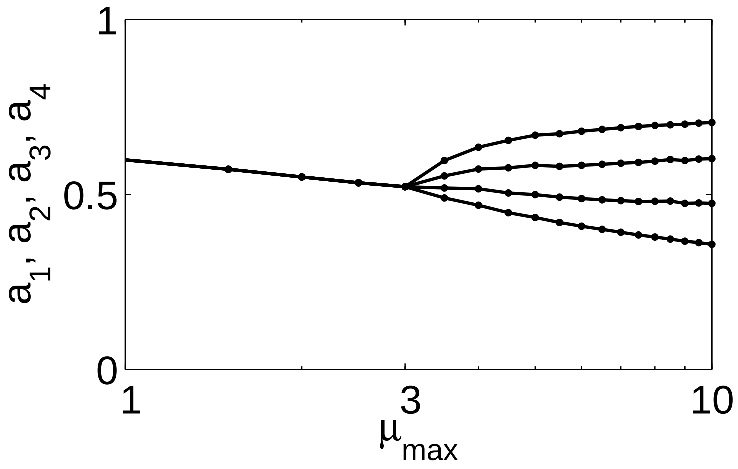

Figure 8.4. Bifurcation diagram with logarithmic μ_{max}-axis that shows the parameters $a_1, \ldots, a_4$ of the optimal gain function within the class $\mathcal{S}_5$. A clear phase transition from the binary step function to a staircase function that uses all available quantization levels occurs at $\mu_{max} \approx 3$. Up to the phase transition the graphs of $a_1, \ldots, a_4$ are in precise agreement with Eq. 8.7.

$\lambda \geq 2$ quantization levels

$$\begin{aligned} \mathcal{S}_\lambda \equiv \quad & \Big\{ g^{\lambda-stair}_{a_1,\ldots,a_{\lambda-1},b_1,\ldots b_{\lambda-2}}(z) : \\ & g^{\lambda-stair}_{a_1,\ldots,a_{\lambda-1},b_1,\ldots b_{\lambda-2}}(z) = b_0 + \textstyle\sum_{l=1}^{\lambda-1}(b_l - b_{l-1})\Theta(x - a_l)\,, \\ & a_1, \ldots, a_{\lambda-1} \in [0,1] \quad \text{and} \quad b_1, \ldots b_{\lambda-2} \in [g_{min}, g_{max}] \Big\}\ , \end{aligned} \tag{8.11}$$

where $b_0 = g_{min}$ and $b_{\lambda-1} = g_{max}$. All these classes together build up a hierarchy of genuine subsets:

$$\mathcal{S}_2 \subset \mathcal{S}_3 \subset \mathcal{S}_4 \subset \ldots \quad . \tag{8.12}$$

and contain the optimal binary step functions given by Eq. 8.6 as a special case. Note, that for $\lambda = 2$ the general notation above might be misleading, because then b_1 does not constitute a free parameter anymore so that we have $g^{2-stair}_{a_1}(z) = b_0 + (b_1 - b_0)\Theta(x - a_1)$. The MMSE for all of these gain functions reads

$$\chi^2[g^{\lambda-const}_{a_1,\ldots,a_{\lambda-1},b_1,\ldots b_{\lambda-2}}] = \frac{1}{3} - \frac{1}{4}\sum_{k=0}^{\infty}\frac{1}{k!}\frac{\left[\sum_{i=1}^{\lambda}\left(a_i^2 - a_{i-1}^2\right)b_{i-1}^k e^{-b_{i-1}}\right]^2}{\sum_{j=1}^{\lambda}\left(a_j - a_{j-1}\right)b_{j-1}^k e^{-b_{j-1}}} \tag{8.13}$$

where $a_0 = 0$ and $a_\lambda = 1$ are the left and right boundaries of the interval, respectively.

In the case of $g_{min} = 0$ and $\lambda = 3, 4, 5$, we evaluated the optimal parameters $a_1, \ldots, a_{\lambda-1}, b_1, \ldots b_{\lambda-2}$ as a function of μ_{max} finding a phase transition at $\mu^c_{max} \gtrsim$

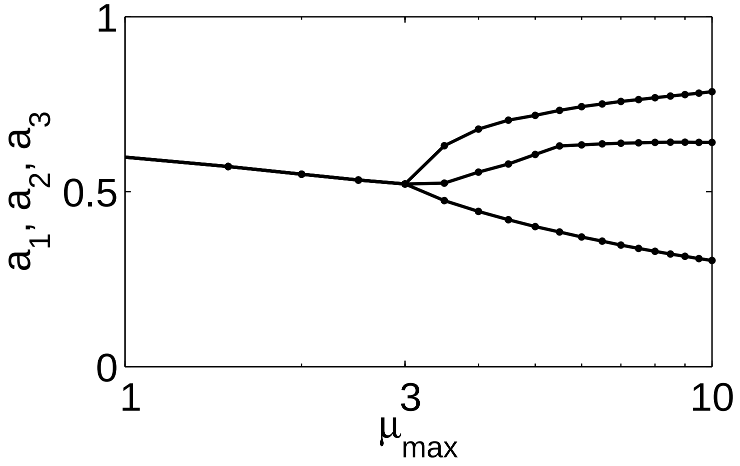

Figure 8.5. Bifurcation diagram with logarithmic μ_{max}-axis that shows the parameters $a_1, \dots, a_3$ of the optimal gain function within the class $\mathcal{L}_3$. A clear phase transition from the binary step function to a piecewise linear function occurs at $\mu_{max} \approx 3$. Up to the phase transition the graphs of $a_1, \dots, a_3$ are in precise agreement with Eq. 8.7.

2.95 (Fig. 8.4). For $\mu_{max} < \mu^c_{max}$ the optimal gain function within $\mathcal{S}_5$ is equal to the optimal step function defined by Eq. 8.6 and Eq. 8.7. For $\mu_{max} > \mu^c_{max}$, however, it makes use of all available quantization levels.

In order to check, whether the binary coding for $\mu_{max} < \mu^c_{max}$ is generically optimal or whether this is rather due to the specific parameterization of $\mathcal{S}_5$, we also considered another function class $\mathcal{L}_\lambda$ that consists of piecewise linear gain functions:

$$g^{\lambda-linear}_{a_1,\dots,a_\lambda,b_2,\dots b_{\lambda-1}}(z) = \begin{cases} b_1 & , \quad 0 < x < a_1 \\ b_1 + (b_2 - b_1)\frac{x-a_1}{a_2-a_1} & , \quad a_1 < x < a_2 \\ b_2 + (b_3 - b_2)\frac{x-a_2}{a_3-a_2} & , \quad a_2 < x < a_3 \\ \vdots & , \quad \vdots \\ b_{\lambda-1} + (b_\lambda - b_{\lambda-1})\frac{x-a_{\lambda-1}}{a_\lambda-a_{\lambda-1}} & , \quad a_{\lambda-1} < x < a_\lambda \\ b_\lambda & , \quad a_\lambda < x < 1 \end{cases} , \qquad (8.14)$$

where $b_1 = g_{min}$ and $b_\lambda = g_{max}$ and it holds $\mathcal{S}_\lambda \subset \mathcal{L}_{\lambda+1}$.

We determined the optimal gain function within $\mathcal{L}_3$, for which the MMSE is given by

$$\chi^2[g^{\lambda-linear}_{a_1,\dots,a_\lambda,b_2,\dots b_{\lambda-1}}] = \frac{1}{3} - \sum_{k=0}^{\infty} \frac{\mathcal{A}^2(k)}{\mathcal{B}(k)} \qquad (8.15)$$

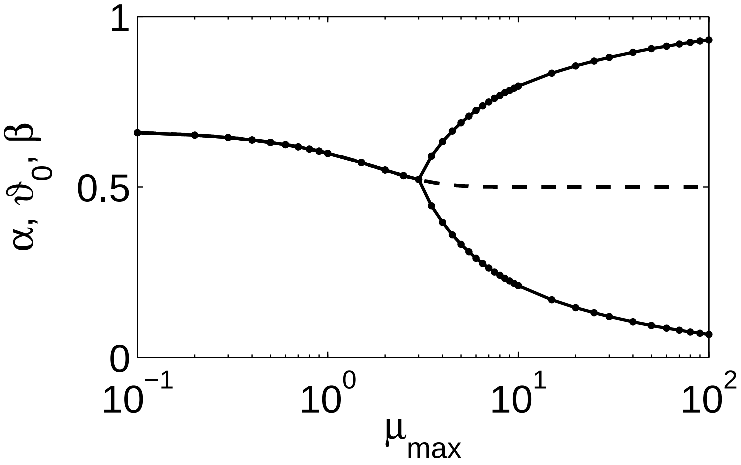

Figure 8.6. Bifurcation diagram with logarithmic μ_{max}-axis that shows the parameters α, β of the optimal gain function within the class $\mathcal{R}_2$. A clear phase transition from the binary step function to a parabolic ramp function occurs at $\mu_{max} \approx 3$. Up to the phase transition the graphs of α, β are in precise agreement with Eq. 8.7. After the phase transition, the width of the parabolic region is permanently increasing. The continuation of the graph of the optimal threshold ϑ_0 is indicated by the dashed line.

where

$$\begin{aligned}
\mathcal{A}(k) &= \frac{a_1^2}{2}\delta_{k,0} + A_{a_1,a_2,0,b_2} + A_{a_2,a_3,b_2,\mu_{max}} + \frac{(1-a_3^2)}{2}\frac{\mu_{max}^k e^{-\mu_{max}}}{k!} && (8.16)\\
\mathcal{B}(k) &= a_1\delta_{k,0} + B_{a_1,a_2,0,b_2} + B_{a_2,a_3,b_2,\mu_{max}} + (1-a_3)\frac{\mu_{max}^k e^{-\mu_{max}}}{k!} && (8.17)\\
A_{\alpha,\beta,\gamma,\zeta} &= \frac{(\beta-\alpha)^2\Gamma_{\gamma,\zeta}(k+2)}{k!(\zeta-\gamma)^2} + \left(\alpha - \frac{\gamma(\beta-\alpha)}{(\zeta-\gamma)}\right) B_{\alpha,\beta,\gamma,\zeta} && (8.18)\\
B_{\alpha,\beta,\gamma,\zeta} &= \frac{1}{k!}\frac{(\beta-\alpha)}{(\zeta-\gamma)}\Gamma_{\gamma,\zeta}(k+1) \quad . && (8.19)
\end{aligned}$$

The corresponding bifurcation diagram for the optimal parameters as a function of μ_{max} is shown in Fig. 8.5, which again exhibits a phase transition at $\mu_{max}^c \approx 3$.

Finally, we consider the function class $\mathcal{R}_2$, which has only two free parameters $\alpha \leq \beta \in [0,1]$ and contains $\mathcal{S}_2$ as well as the asymptotic optimal parabolic function as special cases. The parameterization

$$g_{\alpha,\beta}^{ramp}(z) = \begin{cases} g_{min} & , \quad 0 < x < \alpha \\ ((\sqrt{g_{max}} - \sqrt{g_{min}})\frac{x-\alpha}{\beta-\alpha} + \sqrt{g_{min}})^2 & , \quad \alpha < x < \beta \\ g_{max} & , \quad \beta < x < 1 \end{cases} \qquad (8.20)$$

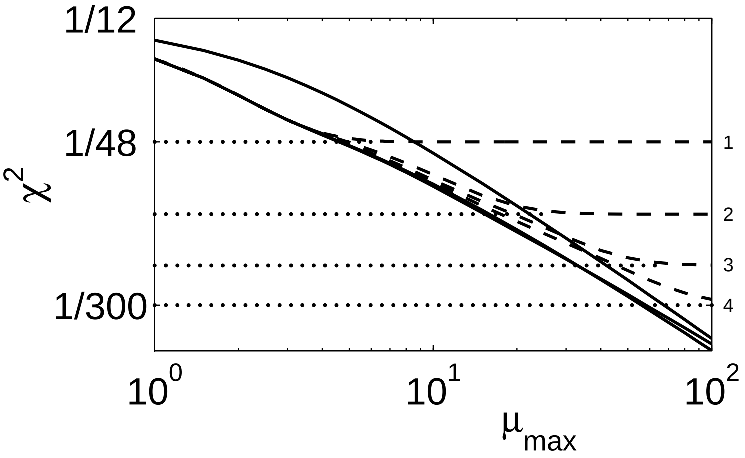

Figure 8.7. Comparison of the MMSEs of the different classes considered in this chapter with the MMSE of the asymptotic optimal gain function g^{asymp} (upper solid). Note the log-log-scale of the axes. The MMSE with respect to $\mathcal{S}_2, \ldots, \mathcal{S}_4$ (dashed) can be distinguished at their index that is printed at their saturation level at the right border of the figure. The MMSEs with respect to L_3 and R_2 can hardly be distinguished (solid, lower), because both graphs are very close to each other over the entire plotted range of μ_{max}.

interpolates between both types of gain functions and their MMSE is given by

$$\chi^2[g^{ramp}_{\alpha,\beta}] = \frac{1}{3} - \frac{1}{2}\sum_{k=0}^{\infty} \frac{1}{k!}\frac{\mathcal{C}^2(k)}{\mathcal{D}(k)} \tag{8.21}$$

where

$$\begin{aligned} \mathcal{C}(k) &= a\frac{b-a}{\sqrt{\mu_{max}}}\Gamma_{0,\mu_{max}}\left(k+\frac{1}{2}\right) + \frac{(b-a)^2}{\mu_{max}}\Gamma_{0,\mu_{max}}(k+1) + \\ &+ (1-b^2)e^{-\mu_{max}}\mu_{max}^k + \delta_{k,0}a^2 \qquad (8.22) \\ \mathcal{D}(k) &= \frac{b-a}{\sqrt{\mu_{max}}}\Gamma_{0,\mu_{max}}\left(k+\frac{1}{2}\right) + 2(1-b)e^{-\mu_{max}}\mu_{max}^k + 2\delta_{k,0}a\,. \qquad (8.23) \end{aligned}$$

Again, we find a phase transition at $\mu_{max}^c \approx 3$ (Fig. 8.6). Taken together the results show that the optimal gain function within $\mathcal{G} = \mathcal{S}_2 \cup \mathcal{S}_3 \cup \mathcal{S}_4 \cup \mathcal{S}_5 \cup \mathcal{L}_3 \cup \mathcal{R}_2$ is the step function $g^{binary} \in \mathcal{S}_2$ given by Eq. 8.6 and Eq. 8.7, provided μ_{max} is smaller than circa three. Due to the diversity of the different considered classes, it is rather unlikely to find a better gain function than the step function, if no such function exists in $\mathcal{G}$.

While all results presented here have been computed with the uniform prior, the existence of the phase transition appears to be independent of the shape of $\rho(z)$. In par-

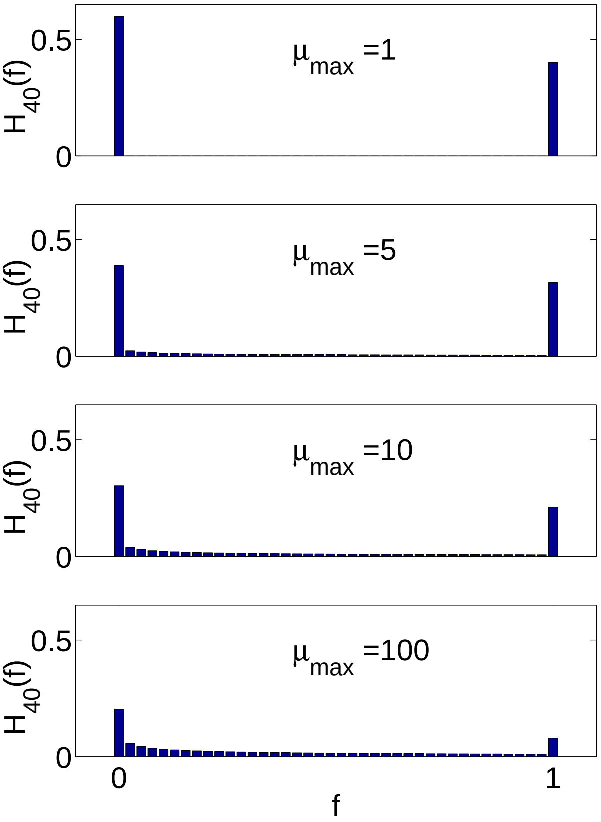

Figure 8.8. The histogram functions of the firing rate distribution induced by g^{ramp} with $\Delta = 40$ bins for different μ_{max} demonstrate that bimodality is a typical feature of optimal firing rate distributions.

ticular, we observed the same qualitative dependence in case of other unimodal distributions (we checked for $\rho(z) = (\nu+1)2^{\nu}(0.5-|0.5-z|)^{\nu}$ with $\nu = 0.2, 0.5, 1, 2, 4, 10$ exhibiting a phase transition at $\bar{\mu}^c = 2.9, 2.9, 2.8, 2.7, 2.6, 2.5$, respectively).

Beyond the phase transition, the optimal encodings of all parameterizations perform similarly well[2] (Fig. 8.7) apart from the regions of saturation in case of $\mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5$ that are due to the bias necessarily induced by piecewise constant gain functions. This bias leads to a lower bound on the MMSE $\chi^2[\mathcal{S}_\lambda] \geq \frac{1}{12\lambda^2}$, which is indicated in Fig. 8.7 by the dotted lines.

Since for $\mu_{max} > \mu^c_{max}$ the MMSE landscape appears to be extremely flat around its minimum, the question for the exact shape of the optimal gain function is rather technical. However, our analysis shows that all optimal encodings exhibit bimodal firing rate distributions with respect to Lebesgue measure. This can be nicely demonstrated with the optimal parabolic ramp function, for which the distribution

[2] We also determined the MMSE with respect to the class $\mathcal{R} := \bigcup_{\gamma\in(0,\infty)} R_\gamma$ in the range $0 \leq \mu_{max} \leq 100$ (not shown). The result was that its MMSE is at most 0.7 % smaller than the MMSE with respect to $\mathcal{R}_2$.

function F of firing rates f is given by

$$F^{ramp}(f) = \begin{cases} 0 & , \quad f < 0 \\ \tilde{\alpha} + (\tilde{\beta} - \tilde{\alpha})\sqrt{f} & , \quad 0 \leq f < 1 \\ 1 & , \quad f \geq 1 \end{cases} \quad , \tag{8.24}$$

where $\tilde{\alpha}, \tilde{\beta}$ are the optimal parameters of the parabolic ramp function, which depend on μ_{max}. In order to illustrate the bimodality of F^{ramp}, we show the corresponding histogram function (see Fig. 8.8), which is defined for any distribution function F by

$$H_\Delta(f) = F(([f/\Delta] + 1)\Delta) - F([f/\Delta]\Delta) \tag{8.25}$$

where $[x]$ denotes the integer part of x and Δ is the bin size. The plotted histogram functions demonstrate that even in the case of $\mu_{max} = 100$ the minimum and the maximum firing rate have a substantially higher probability.

8.5 Analytical study of the phase transition

The finding of a phase transition in all function spaces considered above suggests to check whether the existence of a phase transition can be proved analytically. To this end, two further function classes, $\mathcal{A}_1, \mathcal{A}_2$, are introduced in this section. The two classes contain both, the binary gain function as well as the asymptotic optimal parabolic function as special cases. Furthermore, $\mathcal{A}_1$ is a proper subset of $\mathcal{A}_2$. By numerical optimization within $\mathcal{A}_2$ for various $\bar{\mu}$, we found again a clear phase transition from binary to analog encoding at a critical $\bar{\mu}^c$ with $2.9 < \bar{\mu}^c < 3.0$ (Fig. 8.9, upper). Although the critical value depends on the function space within which the optimization is performed, we did not find any gain function with an error smaller than the MMSE of the step function for $\bar{\mu} < 2.9$.

Our interest in $\mathcal{A}_1$ results from the fact that we can analyze the phase transition in this subset analytically, while $\mathcal{A}_2$ is a quite large function space that is likely to sufficiently approximate all relevant gain functions. Altogether $\mathcal{A}_2$ has six free parameters $a \leq b \leq c \in [0, 1]$, $g_{mid} \in (0, g_{max})$, $\alpha, \beta \in [0, \infty)$, and the parameterization of the gain functions is given by

$$g^{\mathcal{S}_2}(z|a, b, c, g_{mid}, \alpha, \beta) = \begin{cases} 0 & , \quad 0 < z < a \\ g_{mid}\left(\frac{z-a}{b-a}\right)^\alpha & , \quad a < z < b \\ g_{mid} + (g_{max} - g_{mid})\left(\frac{z-b}{c-b}\right)^\beta & , \quad b < z < c \\ g_{max} & , \quad c < z < 1 \end{cases} \quad . \tag{8.26}$$

The integrals entering Eq. 8.4 for the MMSE in case of the tuning function $f^{\mathcal{S}_2}$ then read

$$\int_0^1 x\; p(k|x)\,dx = \frac{1}{k!}\;\Bigg\{\;\frac{a^2}{2}\delta_{0,k} + \frac{(b-a)^2\,\Gamma_{0,f_{mid}}\left(k+\frac{2}{\alpha}\right)}{\alpha(\sqrt[\alpha]{f_{mid}})^2} + \frac{a(b-a)\,\Gamma_{0,f_{mid}}\left(k+\frac{1}{\alpha}\right)}{\alpha\sqrt[\alpha]{f_{mid}}}$$
$$+\frac{(c-b)^2\,\Gamma_{f_{mid},f_{max}}\left(k+\frac{2}{\beta}\right)}{\beta(\sqrt[\beta]{f_{max}}-\sqrt[\beta]{f_{mid}})^2}$$
$$+\left(b-\frac{\sqrt[\beta]{f_{mid}}(c-b)}{(\sqrt[\beta]{f_{max}}-\sqrt[\beta]{f_{mid}})}\right)\frac{(c-b)\,\Gamma_{f_{mid},f_{max}}\left(k+\frac{1}{\beta}\right)}{\beta(\sqrt[\beta]{f_{max}}-\sqrt[\beta]{f_{mid}})}$$
$$+\frac{(1-c^2)}{2}f_{max}^k e^{-f_{max}}\Bigg\}$$

$$\int_0^1 p(k|x)\,dx = \frac{1}{k!}\;\Bigg\{\;a\delta_{0,k} + \frac{(b-a)\,\Gamma_{0,f_{mid}}\left(k+\frac{1}{\alpha}\right)}{\alpha\sqrt[\alpha]{f_{mid}}} + \frac{(c-b)\,\Gamma_{f_{mid},f_{max}}\left(k+\frac{1}{\beta}\right)}{\beta(\sqrt[\beta]{f_{max}}-\sqrt[\beta]{f_{mid}})}$$
$$+(1-c)f_{max}^k e^{-f_{max}}\Bigg\}\;.$$

Numerical optimization leads to the minimal MMSE as a function of $\bar{\mu}$ as displayed in Fig. 8.9 (middle).

The parameterization of the gain functions in $\mathcal{A}_1$ is given by

$$g^{\mathcal{S}_1}(z|w,\gamma) = \begin{cases} 0 & ,\quad 0 < z < \vartheta(\bar{\mu}) - w \\ g_{max}\left(\frac{z-\vartheta(\bar{\mu})+w}{2w}\right)^{\gamma} & ,\quad \vartheta(\bar{\mu}) - w < z < \vartheta(\bar{\mu}) + w \\ g_{max} & ,\quad \vartheta(\bar{\mu}) + w < z < 1 \end{cases} \quad , \qquad (8.27)$$

with $w \in [0,1]$ and $\gamma \in [0,\infty)$. The integrals entering Eq. 8.4 for the MMSE in case of the gain function $g^{\mathcal{S}_1}$ read

$$\int_0^1 z\; p(k|z)\,dz = \frac{1}{k!}\;\Bigg\{\;\frac{(\vartheta(\bar{\mu})-w)^2}{2}\delta_{0,k} + \frac{2w(\vartheta(\bar{\mu})-w)\Gamma_{0,g_{max}}\left(k+\frac{1}{\gamma}\right)}{\gamma\sqrt[\gamma]{g_{max}}}$$
$$+\frac{4w^2\Gamma_{0,g_{max}}\left(k+\frac{2}{\gamma}\right)}{\gamma(\sqrt[\gamma]{g_{max}})^2} + \frac{1-(\vartheta(\bar{\mu})+w)^2}{2}g_{max}^k e^{-g_{max}}\Bigg\}$$

$$\int_0^1 p(k|z)\,dz = \frac{1}{k!}\;\Bigg\{\;(\vartheta(\bar{\mu})-w)\,\delta_{0,k} + \frac{2w\Gamma_{0,g_{max}}\left(k+\frac{1}{\gamma}\right)}{\gamma\sqrt[\gamma]{g_{max}}}$$
$$+(1-\vartheta(\bar{\mu})-w)g_{max}^k e^{-g_{max}}\Bigg\}\,.$$

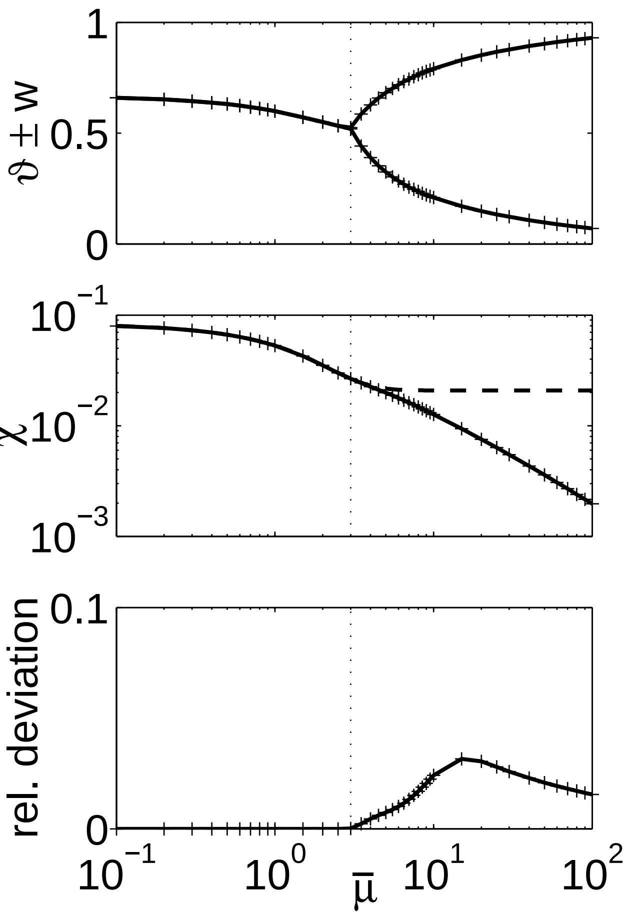

Figure 8.9. The upper panel shows a bifurcation plot for $\vartheta(\bar{\mu}) - w$ and $\vartheta(\bar{\mu}) + w$ of the optimal gain function in $\mathcal{A}_1$ illustrating the phase transition from binary to continuous encoding. The dotted line separates the regions before and after the phase transition in all three panels. Left from this line (i.e. for $\bar{\mu} < \bar{\mu}^c$), the step function given by Eqs. 3 and 4 is optimal. The middle panel shows the MMSE of this step function (dashed) and of the optimal tuning function in $\mathcal{A}_2$ (solid), which becomes smaller than the first one after the phase transition. The relative deviation between the minimal errors of $\mathcal{A}_1$ and $\mathcal{A}_2$ (i.e. $(\chi^2_{\mathcal{A}_1} - \chi^2_{\mathcal{A}_2})/\chi^2_{\mathcal{A}_2}$) is displayed in the lower panel.

The minimal MMSE for these gain functions is only slightly worse than that for $\mathcal{S}_2$. The relative difference between both is plotted in Fig. 8.9 (lower) showing a maximum deviation of 3.2%. In particular, the relative deviation is extremely small around the phase transition. This comparison suggests that a restriction to $\mathcal{A}_1$, which is a necessary simplification for the following analytical investigation, does not change the qualitative results.

The phase transition from binary to analog encoding corresponds to a structural

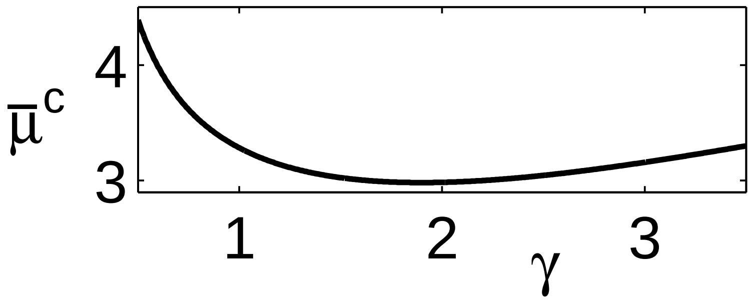

Figure 8.10. The critical maximum mean spike count μ^c is shown as a function of γ (numerical evaluation at $\gamma \in \{0.5, 0.505, 0.51, \ldots, 3.5\}$). The minimum $\mu^c = 2.9857$ at $\gamma = 1.9$ determines the phase transition in $\mathcal{A}_1$.

change of the objective function $\chi^2(w, \gamma)$. In particular, the optimality of binary encoding for $\bar{\mu} < \bar{\mu}^c$ implies that $\chi^2(w, \gamma)$ has a minimum at $w = 0$. The existence of a phase transition implies that with increasing $\bar{\mu}$ this minimum changes into a local maximum at a certain critical point $\bar{\mu} = \bar{\mu}^c$. Therefore, the critical point can be determined by a local expansion of $\chi^2(w, \gamma, \bar{\mu}) - \chi^2(0, \gamma, \bar{\mu}) = \sum_{k=1}^{\infty} g_k(\gamma, \bar{\mu}) \frac{w^k}{k!}$ around $w = 0$, because the sign of its leading coefficient $A_\gamma(\bar{\mu})$ (i.e. the coefficient g_k with minimal k that does not vanish identically) determines, whether $\chi^2(w, \gamma, \bar{\mu})$ has a local minimum or maximum at $w = 0$. Accordingly, the critical point is given as the solution of $A_\gamma(\bar{\mu}) = 0$.

With quite a bit of efforts one can prove that the first derivative of $\chi^2(w, \gamma, \bar{\mu})$ vanishes for all $\bar{\mu}$. The second derivative, however, is a decreasing function of $\bar{\mu}$ and hence constitutes the wanted leading coefficient

$$\begin{aligned}
A_\gamma(\bar{\mu}) &= \frac{1}{4(e^{\bar{\mu}}-1)^2} \Big\{ 8 - 7e^{\bar{\mu}} + 16e^{2\bar{\mu}} + e^{3\bar{\mu}} - \sqrt{1+8e^{-\bar{\mu}}}\left(2 + e^{\bar{\mu}}\left(-3 + e^{\bar{\mu}}(6 + e^{\bar{\mu}})\right)\right) \\
&+ \left(16e^{\bar{\mu}} - 48e^{2\bar{\mu}} - 4e^{3\bar{\mu}} + \sqrt{1+8e^{-\bar{\mu}}}\left(4e^{\bar{\mu}} - 8\left(4 + e^{\bar{\mu}}\right)\right)\right) \frac{\bar{\mu}^{-\frac{1}{\gamma}}}{\gamma} \Gamma_{0,\bar{\mu}}\left(\frac{1}{\gamma}\right) \\
&+ \left(8e^{2\bar{\mu}} + 2\left(5 - 3\sqrt{1+8e^{-\bar{\mu}}}\right) e^{3\bar{\mu}}\right) \frac{\bar{\mu}^{-\frac{2}{\gamma}}}{\gamma^2} \Gamma^2_{0,\bar{\mu}}\left(\frac{1}{\gamma}\right) \qquad (8.28) \\
&- 16e^{\bar{\mu}}\left(e^{\bar{\mu}} - 1\right)\left(\sqrt{1+8e^{-\bar{\mu}}} - 3\right) \frac{\bar{\mu}^{-\frac{2}{\gamma}}}{\gamma} \Gamma_{0,\bar{\mu}}\left(\frac{2}{\gamma}\right) \\
&+ 2e^{2\bar{\mu}}\left(e^{\bar{\mu}} - 1\right)\left(\sqrt{1+8e^{-\bar{\mu}}} - 3\right) \frac{\bar{\mu}^{-\frac{2}{\gamma}}}{\gamma^2} \int_0^{\bar{\mu}} e^{-s} s^{\frac{1-\gamma}{\gamma}} \left(1 - \frac{s}{\bar{\mu}}\right)^{-\frac{1}{\gamma}} \Gamma_{0,\bar{\mu}-s}\left(\frac{1}{\gamma}\right) ds \Big\} \quad .
\end{aligned}$$

Obviously, it is not possible to write the zeros of $A_\gamma(\bar{\mu})$ in a closed form. The numerical evaluation of the critical point $\bar{\mu}^c(\gamma)$ as a function of γ is displayed in Fig. 8.10. Note, that we have treated γ as a fixed parameter, which means that we

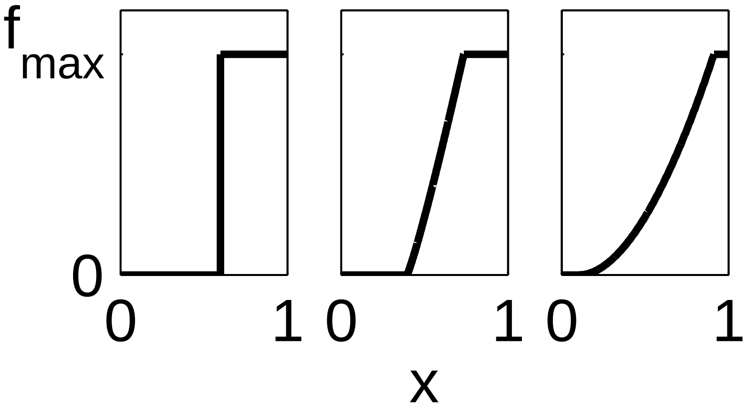

Figure 8.11. Examples of the optimal gain function within $\mathcal{A}_1$ for $\bar{\mu} < \bar{\mu}^c$ (left), $\bar{\mu} = 5$ (middle), and $\bar{\mu} = 50$ (right).

determine the critical point of the phase transition in all subsets $\mathcal{A}_1(\gamma)$ of $\mathcal{A}_1$ that correspond to a fixed γ. It is straightforward to show that the critical point $\bar{\mu}^c$, with respect to the entire class $\mathcal{A}_1$, is given by the minimum of $\bar{\mu}^c(\gamma)$. We determined this value up to a precision of ± 0.0001 to be $\bar{\mu}^c = 2.9857$. The shape of the optimal tuning function before and after the phase transition is displayed in Fig. 8.11. Since the structural change of the objective function is governed by a change of the sign of the leading coefficient, we thus have found a second-order phase transition.

8.6 Discussion

In this chapter, we have derived the shape of optimal gain functions for rate coding, depending on the maximum number of spikes μ_{max} that can be integrated by subsequent neurons. It has been shown that optimal tuning is strictly binary for $\mu_{max} \lessapprox 3$ and $g_{min} = 0$. For larger μ_{max}, optimal gain functions may have regions of analog encoding. However, these gain functions still cause bimodal firing rate distributions at least up to $\mu_{max} = 100$. Within the function class $\mathcal{A}_1$ the phase transition from binary to continuous encoding has been treated analytically. From the point of view of coding efficiency, I therefore conclude that the idea of analog rate signaling is unlikely to be relevant for cortical information processing.

Intuitively, this result is quite conceivable by recognizing that the quantized nature of spike counts imposes already a strong constraint on rate signaling on its own. In this way, it has been argued before by Softky and Koch in [SK93] that the irregular firing of cortical neurons *in vivo* contradicts the possibility of rate coding. In fact, it is rather obvious that the maximum number of spikes k_{max} that can be taken into account by subsequent neurons is limited by their integration time. In other words, a more intuitive way to point out the constraint on analog coding is to describe a rate code as a discrete code with $k_{max}+1$ different symbols. k_{max} has to be small, because the high degree of irregularity in spike timing implies that the effective integration time of cortical neurons is small in comparison with the average interspike interval. Therefore, the mutual information I between the received number of spikes k and the underlying firing rate f has to be small as well, because it cannot be larger than the log index of the symbol set (i.e. $I \leq \log_2(k_{max}+1)$ [CT91]).

Similarly to mutual information, the MMSE χ^2 is bounded by k_{max} as well. For any uniformly distributed signal with variance v, it holds $\chi^2 \geq v/(k_{max}+1)^2$, where the r.h. side corresponds to the mean squared error of the uniform quantizer [GG92]. Furthermore, the two bounds can only be attained in case of noiseless signal transmission.

On the other hand, there is a large body of literature that considers populations of thousands of neurons as the basic units of information processing. This view is motivated by local similarities in the tuning of different neurons as they can be found for instance in a cortical column. One of the first theoretical papers that has established this point of view is [WC72], analyzing the question of how to average the activity of individual neurons, in order to obtain the population rate dynamics. Another early work [GM64] introduced the idea that the balance of excitation and inhibition may provide a source of noise that could explain the high degree of irregularity in the firing of cortical neurons. Combining these ideas, a more recent work [SN98] concludes that noisy firing is used in order to enable analog population rate coding over a large dynamic range.

From the point of view of coding efficiency, however, we have seen in chapter 6 that the strategy of using intensity coding is particularly disadvantageous, if it is used to represent a single analog value by the total population rate: the more neurons encode for the same analog value, the larger becomes the range of μ_{max} , within which intensity coding is not relevant and the larger is the advantage of label-pattern coding over intensity coding. This fact clearly enfeebles the argument presented in [SN98] that it is the goal to achieve a large dynamic range, which explains the high degree of irregularity. Moreover, it is difficult to find a good reason, why cortical processing should not make use of label-pattern coding, which leads to substantially higher precision. In fact, there is experimental evidence that the redundancy in the responses between different neurons is small, supporting the idea of label-pattern coding [RMV01].

Finally, some experimental data shall be discussed, supporting the relevance of binary coding in natural neuronal systems. For example, the *just noticeable difference* (*jnd*) in orientation is known to be almost independent of stimulus contrast over a large range of contrasts [SBS$^+$87]. Together with the fact that the firing rate of most cells in the visual system is strongly affected by stimulus contrast, this sug-

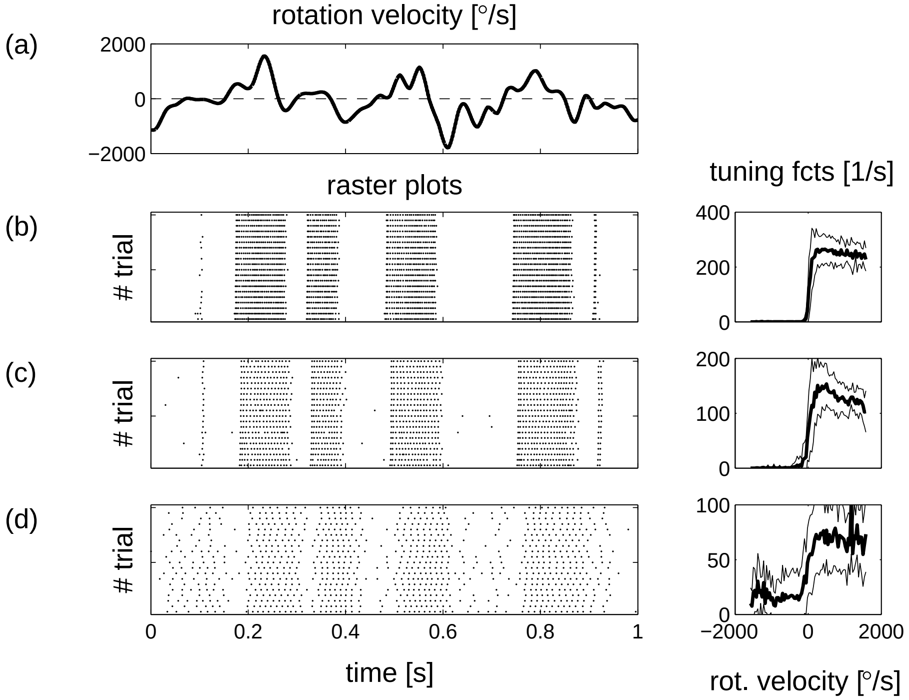

Figure 8.12. Responses of an H1 neuron from an experiment explained in detail in [LBdRvS01] as an example for which the rate depends on the stimulus in a rather binary way. **(a)** The rotation velocity, derived from the yaw component of an actual free flight path (1 s of the total 5 s sequence is shown). **(b,c,d) Left:** raster plots showing a subset of repeated spike train recordings from H1 responding to the same rotation signal shown in (a). **Right:** Corresponding tuning functions (thick) obtained from PSTH with 1 ms bin precision, shifted in time according to the the peak of the cross correlation function (the same kind of regression was used in [BBdRvS00]). Thin lines are given by the mean plus/minus one standard deviation. The difference between (b), (c), and (d) is due to different day times and hence, different light conditions of the recordings. The photon rate per photoreceptor at zenith is 3×10^6 photons/s, 2×10^5 photons/s, and 3×10^2 photons/s, respectively.

gests that the jnd is almost independent of the graded firing rate as well. Vogels [Vog90] recognized the difficulty to explain this finding on the basis of the population vector method. Within the framework of binary coding, however, subsequent neuronal readout should clearly be binary as well. Therefore, the concept of binary coding does not rely on any corrective assumptions as they have been suggested in [Vog90], but on the contrary, the observed contrast independence of the jnd is a generic prediction of binary coding.

The second example shall demonstrate that not all measured tuning functions are smooth. In fact, the H1 neuron of the blow fly constitutes a striking example, for which the firing rate depends on the angular velocity of horizontal rotations of the fly in a rather binary way (see Fig. 8.12, I am grateful to Rob de Ruyter van Steveninck who made the data available to me[3]).

Furthermore, this example illustrates why the hegemony of smooth tuning functions in experimental studies is not a critical finding, but is rather to be expected, even if the encoding is actually binary: In case of the H1 neuron, the binary nature of the response is clearly visible only under optimal light conditions. The smaller the illumination of the stimulus the more unreliable the rate response. This suggests that the tuning is smoothened in the latter case due to stochastic threshold linearization.

In order to explain this point a little more, imagine that the relevant quantity z is not identical with the chosen stimulus parameter s, but it holds $z = s + c$, where c is noise (i.e. an arbitrary contextual quantity that is not under the control of the experimentalist). The variability of c during measurement would lead to a smoothened version $f(s)$ of the actual gain function $f(z)$.

A recent work by Liam Paninski (personal communication) provides yet another example, supporting the point of view that smoothness in measured tuning functions is an expectable artefact rather than a meaningful finding. Using data recorded from a multiple-electrode array implanted in the primary motor cortex of macaques, performing a visually guided manual target tracking task [PFHD03], he showed that tuning functions for manual reaching experiments look much more nonlinear if they are plotted as a function of the 'principal component' instead of the commonly chosen projections onto position or velocity.

The fact that the response of the H1 neuron is the more binary the *better* the light conditions, even suggests an explicit interpretation for the role of intermediate firing rates: instead of permanently sacrificing temporal precision for resolving graded rate differences in order to enlarge the range of distinguishable stimulus parameter values, intermediate rates could represent a state of uncertainty. In fact, a low pass average over a binary random variable can be interpreted as a maximum likelihood estimate of its probability distribution. In this way, a binary code provides a simple solution to the important problem of how to represent uncertain knowledge. Furthermore, in order to obtain a Bayes optimal estimate (w.r.t. 0-1-loss) from such a representation, nothing more is required than a threshold operation. This allows subsequent neurons to choose individually the temporal precision with which they filter the neuronal spike trains depending on the respective required tem-

[3] While we here focus only on that information of the H1 response that can be read out by time-invariant filtering (i.e. 'rate coding'), the authors of the experimental data [LBdRvS01] presented evidence for additional information in the correlation between spikes (cf. [SKdRvSB98; BSK$^+$00]) that might also result from analog changes of the velocity signal. In order to test directly, to which extent the analog signal actually affects the H1 response, I propose to compare the information rate $I[\dot{\phi}(t)]$ of the H1 response in case of an analog input signal $\dot{\phi}(t)$ with the information rate $I[\mathrm{sgn}(\dot{\phi}(t))]$ that is obtained when the analog input signal is replaced with a random telegraph signal such that the sign of both matches for all t. This experiment is presently under way (de Ruyter van Steveninck, personal communication).

poral precision: the less temporal precision is necessary the more precise becomes the representation of the probabilities. Since few samples are virtually sufficient to determine a Bernoulli distribution, this solution of representing uncertainty 'on demand' appears to match the constraints and demands of sensory processing with spiking neurons. In conclusion, such a "*Bernoulli code*" is not only efficient in terms of signal transmission as shown here, but additionally it carries the information in a highly usable form. In this way, the information can be flexibly read out, adjusted to the required temporal precision, with low computational complexity.

8.7 Appendix

Derivation of Eq. 8.7 and Eq. 8.8

We determine the MMSE for the gain function

$$g(z) = g_{max}\,\Theta\left(z-\vartheta\right)) \ , \tag{8.29}$$

which parameterizes $\mathcal{S}_2$ in case for $g_{min}=0$ as a function of μ_{max} and θ. In order to simplify calculations, we substitute $z = x - \frac{1}{2}$ in the following so that Eq. 8.3 is given by

$$p(k|\mu(z)) = \begin{cases} \delta_{0,k} & , \quad -\frac{1}{2} < z < \vartheta - \frac{1}{2} \\ \frac{(\mu_{max})^k}{k!} e^{-\mu_{max}} & , \quad \vartheta - \frac{1}{2} < z < \frac{1}{2} \end{cases} \tag{8.30}$$

and Eq. 8.4 becomes

$$\chi^2[\mu(z)] = E[z^2] - E[\hat{z}^2] = \frac{1}{12} - \sum_{k=0}^{\infty} \frac{\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} z\, p(k|\mu(z))\, dz\right)^2}{\int_{-\frac{1}{2}}^{\frac{1}{2}} p(k|\mu(z))\, dz} \ , \tag{8.31}$$

For the integrals we obtain

$$\begin{aligned} \int_{-\frac{1}{2}}^{\frac{1}{2}} z\, p(k|\mu(z))\, dz &= \frac{1}{k!}\left(\int_{-\frac{1}{2}}^{\vartheta-\frac{1}{2}} z\, \delta_{0,k}\, dz + \int_{\vartheta-\frac{1}{2}}^{\frac{1}{2}} z\, \mu_{max}^k\, e^{-\mu_{max}} dz\right) \\ &= \frac{\vartheta^2-\vartheta}{2}\left(\delta_{0,k} - \frac{1}{k!}\mu_{max}^k\, e^{-\mu_{max}}\right) \end{aligned} \tag{8.32}$$

and

$$\begin{aligned} \int_{-\frac{1}{2}}^{\frac{1}{2}} p(k|\mu(z))\, dz &= \frac{1}{k!}\left(\int_{-\frac{1}{2}}^{\vartheta-\frac{1}{2}} \delta_{0,k}\, dz + \int_{\vartheta-\frac{1}{2}}^{\frac{1}{2}} \mu_{max}^k\, e^{-\mu_{max}} dz\right) \\ &= \vartheta\, \delta_{0,k} + (1-\vartheta)\frac{1}{k!}\mu_{max}^k e^{-\mu_{max}} \quad . \end{aligned} \tag{8.33}$$

Inserting Eq. 8.32 and Eq. 8.33 into Eq. 8.31 yields

$$\begin{aligned}\chi^2 &= \frac{1}{12} - \frac{(\vartheta^2-\vartheta)^2}{4}\,\frac{(1-e^{-\mu_{max}})^2}{\vartheta + (1-\vartheta)e^{-\mu_{max}}} \\ &\quad -\frac{(\vartheta^2-\vartheta)^2}{4(1-\vartheta)}\sum_{k=1}^{\infty}\frac{1}{k!}\,\mu_{max}^k\,e^{-\mu_{max}} \qquad (8.34)\end{aligned}$$

After some steps one obtains

$$\chi^2 = \frac{1}{12}\left(1 - \frac{3\,\vartheta^2}{[(1-\vartheta)(1-e^{-\mu_{max}})]^{-1}-1}\right) \qquad (8.35)$$

by using the equality $\sum_{k=1}^{\infty}\frac{1}{k!}\mu_{max}^k e^{-\mu_{max}} = 1 - e^{-\mu_{max}}$ The derivative of the r.h. side with respect to ϑ has three different zeros. Together with the conditions that the second derivative should be positive and that $\vartheta \in (0,1)$ Eq. 8.7 remains as the unique minimum of Eq. 8.35.

Part II

Challenging Bernoulli Coding

Chapter 9

Interim Conclusion and Introduction to Part II

The overall conclusion from the previous chapters is the plain insight that analog coding does not suit rate coding at physiologically plausible time scales. This result is rather obvious if one starts with a single cell model right from the beginning. In this thesis, however, it emerged as an optimal solution to the *population coding* problem of representing an N-dimensional vector on the basis of N neurons. Furthermore, this conclusion is backed up by the studies in chapter 6 and 5, which provide a strong argument that the importance of label-pattern coding only increases if one reduces the dimensionality of the vector to be represented. Moreover, the model used to derive this result has not been specifically set up to support this particular hypothesis, but the standard framework of population coding has been used, which by the choice of the squared error loss is rather tailored to the goal of analog encoding.

It is not at all an established point of view that graded rate differences are not a useful means for the design of neuronal rate codes. First of all, there is clearly a rather unspecific preference for smooth functions, which is due to the mathematical experience that singularities of nondifferentiability can be very nasty to deal with. Therefore, the mathematically educated researcher typically feels that neurons should rather avoid the use of such freaky tuning functions. On the other hand, the difficulty to find an example of a measured tuning function that looks binary leads inevitably to the impression that 'empirical evidence' speaks against the idea of binary coding. These intuitive arguments are so fundamental that it is hard to avoid getting affected by them even if one is aware of the fact that it is, for instance, too naive to take the shape of a measured tuning function as evidence against binary coding, as long as there is large variability[1] in the neuronal response.

[1] The variability probably reflects a signal, which is actually relevant for neuronal processing [ASGA96; Fer96; TKGA99].

There are, however, also some more specific objections against the Bernoulli coding hypothesis. In particular, there are serious objections against the experimental support presented above. For instance, Peter Dayan pointed out to me that the finding of an adaptive rescaling of the dynamic range of the H1 neuron's tuning with respect to the variance of the input signal (i.e. the horizontal rotation velocity) [BBdRvS00] suggests that intensity coding is of meaning there. Apart from that, Rob de Ruyter van Steveninck (who clearly knows his own data much better than I do) also does not follow the interpretation given above, but he believes that the information carried by correlations between subsequent spikes are relevant for the neuronal processing in the fly. This kind of neuronal encoding cannot be treated within the rate coding paradigm at all.

In fact, I have to admit that the 'experimental support' given above can at best function as an attempt to weaken the psychological barrier against strongly nonlinear tuning, while until today I have got no idea how it may be feasible to critically test this hypothesis with the means available to date. However, there is still something what we can do on the theoretical side: Since it is a wide-spread belief that for certain functional demands, other than coding efficiency, smooth tuning is important more than accuracy, it is the purpose of the second part of this thesis to depart from the efficiency argument and to seek other criteria to challenge the Bernoulli coding hypothesis.

One such argument is the limited computational power of neuronal read out. In fact, many studies of population codes are based on the population vector method or on the LMS-estimator, which are both not capable to take advantage of label-pattern coding. As has already been shown in chapter 7, however, a very simple threshold nonlinearity, which appears to be more than plausible with respect to the basic biophysical properties of neurons, is indeed sufficient to make use of label-pattern coding.

The following two chapters will address two other intuitions. One intuition is that analog encoding strategies, like population rate coding, are more robust against cell death and other sources of unreliability. This issue will be discussed in the next chapter. Another criterion is the question, whether a coding scheme can actually be realized when taking the neuronal dynamics into account. To this end, chapter 11 investigates the possibility of realizing analog and label-pattern rate coding schemes by means of integrate-and-fire neurons.

Chapter 10

Robustness: Bernoulli vs Population Rate Coding

The need for a neuronal coding scheme that is robust against cell death and the corruption of action potentials seems to support the idea of population rate coding, where the labels of different neurons need not to be distinguished. In order to test this intuition, this chapter investigates the efficiency and robustness of a population rate coding scheme in comparison to a label-pattern coding scheme, using identical noise models. As it turns out, not only the efficiency of population rate coding is substantially worse than that of a label-pattern coding scheme, but the latter also remains to be superior, independent of the number of neurons that are randomly selected to be erased.

A good model for the description of a binary rate code is the *Bernoulli neuron model.* The Bernoulli model can be related to what is measured in a PSTH if the bin width chosen is sufficiently small. Then one will find at most one spike within any given bin in all trials due to the refractory period of neurons. Hence, the average spike count over trials for each bin t will be a number $p(t) \in [0, 1]$, representing the fraction of trials, where a spike occurred.

In contrast to a rate signal transmitted by a single neuron, population rates are not necessarily slow, but with an increasing number of neurons, it becomes possible to get a reliable estimate of $p(t)$ simply by averaging over a population of identically tuned neurons. This idea of population rate coding is, in particular, frequently considered as the biological basis for analog signaling and computation in more abstract neural network models.

10.1 Analysis

Consider a population of N neurons $k = 1, \ldots, N$, whose spike trains together represent a certain real-valued signal $x(t)$ with finite dynamic range. E.g. $x(t)$ could stand for the population rate of all presynaptic neurons [SN98] or for the correlation of two moving objects in the visual field or anything else one might consider relevant for signal processing in the brain. Without restriction of generality, we can choose a scale such that $x \in [0, 1]$ holds. For the sake of simplicity, we focus on temporally uncorrelated discrete time signals $x(t_m)$ with a discretization period $\Delta T = t_m - t_{m-1}$ that is smaller than the neurons minimal interspike interval. In this case, the neuronal response at each instance of time is either $s_k(t_m) = 1$ (spike) or $s_k(t_m) = 0$ (no spike) and the conditional distribution

$$P(s_1(t_m), \ldots, s_N(t_m)|x(t_m)) \tag{10.1}$$

captures all available information about the signal x carried by the spike trains $(s_1, \ldots, s_N)$ (provided the encoding is constant). Assuming that Eq. 10.1 can be written in product form, an encoding is completely specified by N activation functions

$$p_k^{spike}(x) = P(s_k(t_m) = 1|x(t_m)) \tag{10.2}$$

that tell for each neuron $k = 1, \ldots, N$ how much it is driven by a given input x.

While it is not difficult to imagine that a neuron can be reliably prevented from firing, for instance via apical shunting inhibition, there are several reasons to assume that the activation functions can take values only substantially below one: for example, the internal state of the neuron constitutes a source of variability that is independent of the signal $x(t)$, because the excitability of a neuron substantially depends on the time elapsed since it has generated the last spike. Furthermore, the reliability with which a single spike contributes to the entire signal representation is limited by the reliability of synaptic transmission, which can be quantified by the probability of synaptic release p^{rel} [Zad98].

In this model, all possible corruptions of spikes due to unreliable synaptic transmission as well as any other sources of noise are captured by an upper bound for the probabilities of spike generation $p_k^{spike}(x) \leq p^{rel}$.

10.1.1 Rate coding with high noise level

The signal of a population rate coding scheme is given by the population spike count $c(t_m) = \sum_{k=1}^{N} s_k(t_m)$, which is invariant under random permutations of the neuron index k. If p^{rel} is small, the probability mass function $P(c|x)$ for the spike count is well approximated by a Poisson distribution with mean $\mu(x) = \sum_{k=1}^{N} p_k^{spike}(x)$.

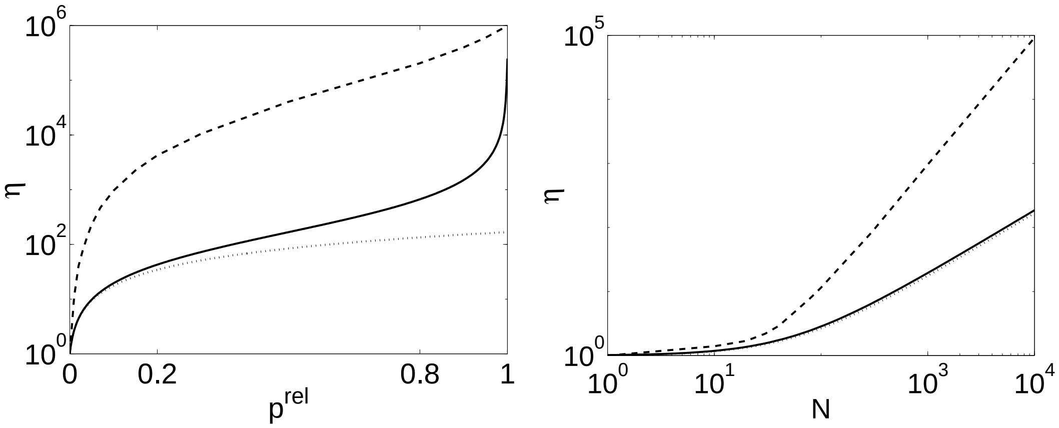

Figure 10.1. Left figure shows the mean square efficiency η as a function of the spike reliability p^{rel} ($N = 1000$) in the case of optimal rate coding (solid), label-pattern coding (dashed), and Poissonian approximation (dotted). **Right** figure shows η as a function of the population size N ($p^{rel} = 0.1$) in the corresponding cases.

Therefore, the population rate coding scheme then is completely specified by an arbitrary function $\mu(x)$ with $0 \leq \mu \leq Np^{rel}$.

The efficiency of the different coding schemes is measured by the ratio

$$\eta = \frac{Var[x]}{E[(x - \hat{x}(\xi))^2]} \leq \frac{Var[x]}{\mathrm{E}\left[Var[x|\xi]\right]} \tag{10.3}$$

where ξ denotes the neuronal response under consideration, and $\hat{x}(\xi)$ is the corresponding estimator used. In case of asymptotic normality, the log of the r.h. side of Eq. 10.3 becomes equal to mutual information. Using $\mu(x) = Np^{rel}x$, this quantity is evaluated and shown in Fig. 10.1 for different coding schemes. On the left it is plotted as a function of p_{rel} in case of $N = 10^3$. On the right it is shown as a function of N for $p^{rel} = 0.1$, which is a typical value of synaptic release probability [HS97].

In case of the Poisson approximation, the efficiency of the intensity coding strategy is given by

$$\eta_{Poisson} = 1 + \frac{Np^{rel}}{6} \tag{10.4}$$

and plotted as dotted line in Fig. 10.1.

10.1.2 Optimal population rate coding

Even in case of population rate coding, the most efficient encoding strategy is again based on label-pattern coding. More specifically, it is optimal, if there is one sub-population denoted by the index set $S_1(x)$, within which all neurons have maximum

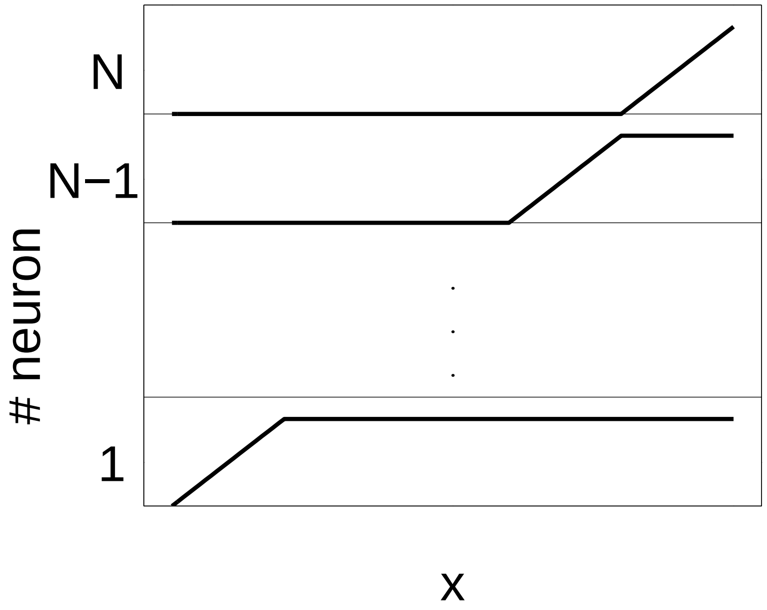

Figure 10.2. The diagram sketches the likelihood functions $p_k^{spike}(x)$ of the population.

probability of spike generation ($p_k^{spike}(x) = p^{rel} \, \forall \, k \in S_1(x)$) and the residual neurons of the complement set $S_0(x)$ have minimum probability ($p_k^{spike}(x) = 0 \, \forall \, k \in S_0(x)$). While the difference to the Poissonian case approaches zero in the limit of small spike probabilities, it becomes large when the boundary p^{rel} is increased. In order to allow for a direct comparison with the Poissonian case, we define the likelihood functions $p_k^{spike}(x)$ such that the mean spike count remains the same as in the previous example (i.e. $\sum_{k=1}^{N} p_k^{spike}(x) = N p^{rel} x$). (see Fig. 10.2). The corresponding result for the mean square efficiency

$$\eta_{binary} = 1 + \frac{N p^{rel}}{6 \frac{N-1}{N}(1 - p^{rel}) + \frac{1}{N}(6 - 4p^{rel})} \tag{10.5}$$

is also displayed in Fig. 10.1 (solid line). This shows that even in case of population rate decoding it is advantageous if the tuning functions are strongly nonlinear.

10.1.3 Label-pattern coding

Now, we ask for the coding accuracy that can be achieved, if subsequent neuronal readout depends on the neuron index as well. A simple estimator that essentially makes use of the neuron index is given by

$$\hat{x}(s_1, \ldots, s_N) := \max \left\{ \frac{1}{N}, \frac{k}{N} \Theta \left(s_k - \frac{1}{2} \right) \middle| \, k = 1, \ldots, N \right\} . \tag{10.6}$$

The resulting efficency is displayed in Fig. 10.1 by the dashed line. A remarkable feature of this label-pattern coding scheme is that its efficency scales with N^2 in contrast to the efficency of the optimal rate coding schemes, which is proportional to N only (right). Furthermore, it is important to note that this superiority of the label-pattern coding scheme is not achieved at the cost of a reduced robustness: the advantage of label-pattern coding against rate coding holds for a given number of neurons even if p^{rel} becomes very small (left). Furthermore, the worst case error $W = \max_x E[(x-\hat{x})^2|x]$ (not shown) is close to the mean squared error $E[(x-\hat{x})^2]$.

If neurons are randomly erased, the distribution of the thresholds of the neurons looses its regularity. In the worst case the thresholds then become uniformly distributed over the interval. In that case, however, the error increases at most by a factor 2, namely if $p_{rel} = 1$. If $p_{rel} < 1$, the error increases even less. In Fig. 10.1 a decrease by a factor of two corresponds to an almost invisible shift of the dashed line. In conclusion, the superiority of the label-pattern coding scheme holds true also if one takes random cell death into account.

10.2 Appendix

10.2.1 Derivation of Eq. 10.4

With $x \sim U(0,1)$ and $\mu(x) = Np^{rel}x$ it follows

$$\begin{aligned}
\mathrm{E}[x] &= \frac{1}{2} && (10.7)\\
\mathrm{E}[k] &= \mathrm{E}[\mu(x)] = Np^{rel}\int_0^1 x\,dx = \frac{Np^{rel}}{2} && (10.8)\\
\mathrm{E}[xk] &= \mathrm{E}[x\mu(x)] = Np^{rel}\int_0^1 x^2\,dx = \frac{Np^{rel}}{3} && (10.9)
\end{aligned}$$

and consequently the cross-covariance reads

$$C_{xk} = \mathrm{E}[xk] - \mathrm{E}[x]\,\mathrm{E}[k] = \frac{Np^{rel}}{12} \quad . \qquad (10.10)$$

Because of $\mathrm{Var}[k|\,x] = \mathrm{E}[k|\,x] = \mu(x)$ for the Poisson model, it holds

$$\begin{aligned}
\mathrm{Var}[\mathrm{E}[k|\,x]] &= \mathrm{Var}[\mu(x)] = (Np^{rel})^2\mathrm{Var}[x] = \frac{(Np^{rel})^2}{12} && (10.11)\\
\mathrm{E}[\mathrm{Var}[k|\,x]] &= \mathrm{E}[\mu(x)] = \frac{Np^{rel}}{2} \quad . && (10.12)
\end{aligned}$$

Since k is 1-dimensional, this determines the auto-covariance

$$C_{kk} = \mathrm{Var}\left[k\right] = \mathrm{Var}\left[\mathrm{E}\left[k|\, x\right]\right] + \mathrm{E}\left[\mathrm{Var}\left[k|\, x\right]\right] = Np^{rel}\left(\frac{Np^{rel}}{12} + \frac{1}{12}\right) \quad . \tag{10.13}$$

In summary, we obtain for the LMMSE

$$\begin{aligned}
\chi^2_{LMS} &= \mathrm{Var}\left[x\right] - \frac{C_{kx}^2}{C_k k} = \frac{1}{12}\left(1 - \frac{Np^{rel}}{12\left(\frac{Np^{rel}}{12} + \frac{1}{2}\right)}\right) \\
&= \frac{1}{12}\left(1 - \frac{Np^{rel}}{Np^{rel} + 6}\right) \\
&= \underline{\underline{\frac{1}{12} \cdot \frac{6}{Np^{rel} + 6}}}
\end{aligned} \tag{10.14}$$

and hence for the efficiency

$$\eta = \frac{1}{12\chi^2_{LMS}} = 1 + \frac{Np^{rel}}{6} \tag{10.15}$$

10.2.2 Derivation of Eq. 10.5

The cross-covariance does not depend on the noise model and hence is given by Eq. 10.10 again.

In case of the optimal population rate encoding strategy, $\lfloor Nx \rfloor$ neurons are activated at maximum (i.e. p^{rel}), $N - \lfloor Nx \rfloor - 1$ neurons have zero probability to fire, and only one neuron has an intermediate probability

$$p_i = \left(x - \frac{\lfloor Nx \rfloor}{N}\right) p^{rel} \quad . \tag{10.16}$$

Therefore, we can conclude that

$$\mathrm{Var}\left[k|\, x\right] = \sum_{j=1}^{N} \mathrm{Var}\left[k_j|\, x\right] = \lfloor Nx \rfloor p^{rel}(1 - p^{rel}) + 0 + p_i(1 - p_i) \quad . \tag{10.17}$$

Integrating over the N intervals $(0, \frac{1}{N})$, $(\frac{1}{N}, \frac{2}{N})$, $\ldots(\frac{N-1}{N}, 1)$ yields

$$\begin{aligned}
\mathrm{E}\left[\mathrm{Var}\left[k|\, x\right]\right] &= \frac{1}{N}\sum_{m=0}^{N-1} mp^{rel}(1 - p^{rel}) + N\int_0^{\frac{1}{N}} Np^{rel}x(1 - Np^{rel}x)\, dx \\
&= \frac{p^{rel}(1 - p^{rel})}{N} \cdot \frac{N(N-1)}{2} + N^2 p^{rel}\frac{1}{2N^2} - N^3(p^{rel})^2\frac{1}{3N^3} \\
&= \frac{N-1}{2}p^{rel}(1 - p^{rel}) + \frac{p^{rel}}{2} - \frac{(p^{rel})^2}{3} \quad .
\end{aligned} \tag{10.18}$$

In summary, we obtain for the LMMSE

$$\begin{aligned}
\chi^2_{LMS} &= \frac{1}{12} - \frac{Np^{rel}}{12\left(Np^{rel} + \frac{12}{Np^{rel}}\left[\frac{N-1}{2}p^{rel}(1-p^{rel}) + \frac{p^{rel}}{2} - \frac{(p^{rel})^2}{3}\right]\right)} \\
&= \frac{1}{12}\left\{1 - \frac{Np^{rel}}{Np^{rel} + 6\frac{N-1}{N}(1-p^{rel}) + \frac{6}{N} - \frac{4p^{rel}}{N}}\right\} \\
&= \underline{\underline{\frac{1}{12}\cdot\frac{6\frac{N-1}{N}(1-p^{rel}) + \frac{1}{N}(6-4p^{rel})}{Np^{rel} + 6\frac{N-1}{N}(1-p^{rel}) + \frac{1}{N}(6-4p^{rel})}}}
\end{aligned} \tag{10.19}$$

Chapter 11

Rapid Population Rate Coding

In this chapter, we will take a large step from the abstract, discrete-time, memoryless encoding models considered above towards more detailed models, which, in particular, take the basic dynamics of neurons into account. In other words, we are departing from merely spatially structured models and ask how the realization of rapid rate codes is possible, depending on the chosen encoding strategy. As it will turn out, the smoothness of tuning has strong implications for the dynamics of neuronal signal transmission.

11.1 Rate coding and the independence of spikes

When I was writing this thesis, I was not sure, whether I should ascribe the discrete-time, memoryless neuron models, used throughout this thesis, to 'rate coding' or whether I should better invent a new category for that right from the beginning. The problem is that the term 'rate code' is used with so many different meanings, so that misunderstandings become highly likely. Nevertheless, for the sake of clarity, I decided to stick with 'rate coding' to begin with and to postpone a more detailed discussion of temporal coding issues to this chapter.

In the ongoing discussions about the neural code, the term '*rate coding*' has mainly been used by proponents of "*temporal coding*" as an equivalent to the notion of a low temporal resolution. While it is true that many experimental studies consider spike counts obtained from time windows on the order of seconds, I think, it is not very useful to assign the term 'rate coding' to a particular time scale. I believe, in particular, that everyone would agree that the relevant time scale has to be smaller than, say 50 ms, because any larger time scale is actually not compatible with the speed of neuronal processing.

“Temporal coding” in turn, actually sounds like a much stronger claim than standing for nothing more than the quite obvious doubts regarding the slow time scale. In fact most proponents of “temporal coding” place the objections against “slow rate coding” not with respect to the issue of temporal precision *per se*, but in order to support more specific ideas like e.g. that the **relative** spike timing (i.e. the interspike interval for instance) is used in the brain to distinguish between different stimulus features. This belief, however, is by far more arguable than a high temporal precision. I quote [DA01]:

> *When careful studies have been done, it has been found that some information is carried by correlations between two or more spikes, but this information is rarely larger than 10% of the information carried by spikes considered independently. Of course, it is possible that, due to our ignorance of the “real” neural code, we have not yet uncovered or examined the types of correlations that are most significant for neural coding. Although this is not impossible, we view it as unlikely and feel that the evidence for independent-spike coding, at least as a fairly accurate approximation, is quite convincing.*

In the following, I will refer to this point of view as the *independent-spike hypothesis*. Another way to put it is to say that the Poisson model is sufficient to describe the neuronal response to the stimuli that are processed in the neural networks of the brain. The independent-spike hypothesis is helpful, when giving a more comprehensive overview on ’temporal coding’, to point out that it has been related to the following different issues:

- As the opposite to slow rate coding (in this case, “temporal coding” means nothing more than a high temporal precision, which is compatible with the independent-spike hypothesis)
- As the encoding of static stimulus features in the temporal structure of the neuronal response (i.e. if a stimulus signal is presented with a cut off frequency ν and the unfiltered neuronal response contains more information about the stimulus signal than a low pass filtered version of the response with cut off ν [TM95]). Again, this definition is compatible with the independent-spike hypothesis.
- As a spike-pattern code, which makes distinctive use of serial correlations between spikes, standing in opposition to the independent-spike hypothesis.

Let us now consider as to how the discrete-time memoryless channel is related to these three issues. The first point is easy. As mentioned several times before, the psychophysical performance as well as the effective integration time of neurons *in* *vivo* provides strong evidence that the relevant time scale is short, say in the range of 1-50 ms.

The last point is more difficult to answer, because it is not clear when the use of serial correlations between spikes is ’distinctive’, and when it is not. Strictly speaking, the independent-spike hypothesis is fulfilled, if and *only if* the neurons function as Poisson spike generators. Strictly speaking, however, neurons cannot realize a Poisson process for principle reasons. The finite temporal extension of a spike and, in particular, the subsequent refractory period necessarily generate serial correlations. Therefore, the independent-spike hypothesis has to be understood as an approximation. This, however, requires to define, where to draw the line between relevant and irrelevant correlations.

The usual practice, to refer to the estimated fraction of the amount of information at this point, is dangerous because the notion of a hypothesis should not rely on data. The Bernoulli coding hypothesis proposed in this thesis can, in particular, take substantial advantage of serial correlations between spikes, but it works with a pure Poisson process as well. In order to make this point clear, it is instructive, to seek a description of the discrete-time memoryless channel considered above in terms of a point process. To this end, let us assume a bin width $T = 5$ ms, within which the signal to be represented is taken as constant. In case of a Poisson noise model, it is straightforward to define a corresponding Poisson process. Its piecewise constant intensity function directly reflects the firing rates given by the tuning functions.

As in the previous chapter, however, we also considered the Bernoulli noise model, which in a sense is more realistic than the Poisson noise model, because real neurons cannot generate more than one spike within a sufficiently small time bin. While for $p^{rel} \to 0$ the difference to the Poisson model vanishes, the Bernoulli model is very different for large p^{rel}. In order to implement such a code as a point process, one has to resort to a renewal process, and it is impossible to approximate this situation by a Poisson process. Clearly, the reason for this problem is that a small spike count variability can be achieved only if the intensity function of a point process may depend on the last spike. In other words, spike correlations are definitely relevant in this case and increase the mutual information dramatically, but nevertheless, the *signal* can be read out “instantaneously” and is clearly not related to interspike intervals or more complex spike patterns.

I think, this is a striking example of how misunderstandings can arise, if we do not set up specific models, but vaguely refer to ’spike correlations’ only. In fact, the notion of ’spike correlations’ is way too ambiguous. What we need are clear models that hypothesize about what is the *signal* as opposed to the noise (even if this implies the risk that these models may be falsified by data, or, conversely, that available data cannot be used any more to support certain statements).

Experimentally, the question of the statistical independence of spikes is related to the

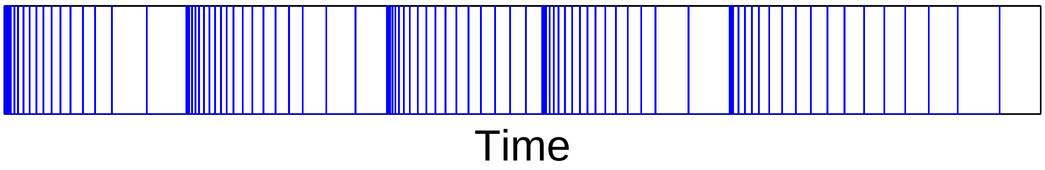

Figure 11.1. Example of spike train with exponential ISI statistic ($C_V = 1$) and strong regularities.

measurement of the irregularity in sequences of interspike intervals and the response variability. The commonly used measure for the irregularity of a spike train is the *coefficient of variation*, which is defined by

$$C_V = \frac{\sqrt{\mathrm{Var}\left[T^{ISI}\right]}}{\mathrm{E}\left[T^{ISI}\right]} \quad . \tag{11.1}$$

The finding that the C_V values for cortical neurons are typically close to one shows that the degree of irregularity is compatible with that of a Poisson process, for which $C_V = 1$ in case of a constant intensity function. Furthermore, the shape of measured ISI distributions also resembles the exponential shape of a homogenous Poisson process[1] in some detail.

For inhomogeneous point processes, however, the finding of an exponentially shaped ISI histogram is not a unique feature of the Poisson process. E.g. it is possible to obtain an exponentially distributed ISI histogram also in case of spike trains that have strong regularities (see Fig. 11.1). Hence, what is actually necessary to consider is the entire set of conditional ISI distributions, which depend on the length of the preceeding interspike interval. Accordingly, a more significant measure of irregularity would be the mean conditional coefficient of variation, which is defined by

$$\overline{C_V} = \mathrm{E}\left[C_V(T^{ISI}_{pre})\right] = \int_0^\infty C_V(T^{ISI}_{pre})\,\rho(T^{ISI}_{pre})\,dT^{ISI}_{pre} \quad . \tag{11.2}$$

In the equation above, $C_V(T^{ISI}_{pre})$ is the coefficient of variation with respect to the conditional ISI distribution that depends on T^{ISI}_{pre} of the preceeding interspike inter-

[1] For a constant intensity function, the exponential shape of the ISI distribution follows directly from Eq. 2.11.

val, i.e.

$$C_V(T_{pre}^{ISI}) = \frac{\sqrt{\mathrm{Var}\left[T^{ISI} \middle| T_{pre}^{ISI}\right]}}{\mathrm{E}\left[T^{ISI} \middle| T_{pre}^{ISI}\right]} \quad . \tag{11.3}$$

Unfortunately, this data analysis has not been carried out so far. Although it is visible by eye that $\overline{C_V}$ will not differ too much from C_V for cortical spike trains, it would still be nice to have a quantitative estimate.

Another way to take the question of statistical independence of successive interspike intervals into account, which has been brought to neurophysiology by Rodieck, Kiang, and Gerstein in 1962 [RKG62], is the joint ISI histogram, where one plot each interspike interval over its preceeding ISI. In case of statistical independence, this diagram has to be symmetric with respect to the diagonal. More precisely, the joint density $\rho(T_{pre}^{ISI}, T^{ISI})$ should equal the product of the marginals, which are both identical to the unconditional ISI density.

However, even if the spike trains matched perfectly all statistical properties of a Poisson process that can be determined without control of the underlying intensity function, this would not necessarily imply the Poisson model to be a valid description of neuronal signal transmission. Because neurons are driven systems, the question whether they instantiate a Poisson process or not may depend on the definition of the input signals. In fact, it is well possible that spike generation considered as a function[2] of the synaptic inputs constitutes a renewal model but nevertheless instantiates a Poisson process with respect to a more abstract input signal of interest, like e.g. a parameter of a visual stimulus.

We illustrate this point with the LIF neuron. In this model, the signatures of statistical independence of successive spikes considered so far can be achieved, e.g. if the input is a Dirac delta comb, for which the delta functions are scaled by V_{th} and temporally distributed with respect to a Poisson process. This shows that the description of spike generation with an LIF model does not uniquely define a particular renewal process, but it still allows for a large set of possible candidates that even includes the Poisson process. Or to put it differently, the LIF neuron may function as a part of a system which instantiates a Poisson model; the only known thing is that it definitely cannot realize a Markov process of higher orders than renewal.

Up to now, we only considered those aspects of statistical dependencies between successive spikes that can be analyzed on the basis of a single spike train. For non-stationary point processes, however, it does not make any sense to refer to a single spike train only, but it is essentially a description of an infinite set of spike trains

[2] In precise mathematical terms it is not a function, but a stochastic operator.

with certain common properties. Hence, it is not sufficient to analyze a single spike train only, but we need to check the variability between different samples drawn from the same process as well. Experimentally, the process as a whole, and the variability between different samples of this process clearly depend on the control conditions used to define operationally different recordings as different samples of the same process.

In neurophysiological studies there are typically two different control conditions used. If the neuronal response to the direct injection of a current signal is recorded, the set of responses that have been triggered by identical input signals are considered as samples of the same process. If the neuronal response to the presentation of exogenous stimuli (say e.g. visual stimuli) is measured, all spike trains that have been triggered by the same stimulus are considered as samples of the same process. Clearly, the second control condition does not imply that the neuron receives the same synaptic input in different trials of the same exogenous stimulus. Therefore, it is important to say, which control condition a neuronal response model refers to, when discussing its statistical properties.

In chapter 2, we have mainly considered the case where the input current is used as control condition. As reported above, this situation is not well described by a Poisson process, but requires to take at least the last spike time into account, as in the LIF model. In particular, the responses to the same input current signal look very similar across different trials, which is not compatible with the Poisson process.

The major reason, why the Poisson model is frequently used to model experimental studies stems from the fact that under the second control condition almost all stimulus features investigated so far constrain the neuronal response so weakly that even the total spike count is not reproducible from trial to trial, but varies in a manner very similar to what one would expect from the Poisson process.

The *spike count variability* is commonly measured by the variance of the counting statistics $P(k|t_1, t_2)$ with respect to the interval (t_1, t_2) divided by its mean

$$F(t_1, t_2) = \frac{\mathrm{Var}\,[k|\,t_1, t_2]}{\mathrm{E}\,[k|\,t_1, t_2]} \tag{11.4}$$

In the limit $t_2 - t_1 \to \infty$ this ratio is called the *Fano factor* F_∞ and it holds $F_\infty = C_V^2$ for all homogeneous renewal processes. In case of the integrate-and-fire models with constant input, both, the coefficient of variation as well as the Fano factor equal zero. In contrast, in case of the homogeneous Poisson process both measures equal one. Since the counting statistics of the inhomogeneous Poisson process does not depend on the choice of the time window, $F(t_1, t_2)$ is always one independent of the time bin (t_1, t_2).

In most experiments, the measured variability between trials controlled by the same visual stimulus is close to or even exceeds that of a Poisson process. While the excess variability can only be explained by changes in uncontrolled parameters of the input, the ubiquitous finding of a large amount of variability leads to the impression that neural encoding (in cortex) of information about sensory data does not make use of knowledge about when previous spikes have been generated. In other words, the large variability is a necessary condition for the *independent-spike hypothesis*, but it is clearly not sufficient.

The issue, how well the independent-spike hypothesis is actually justified, is not a simple matter. In contrast, the irregularity of discharge is well established, because it can be determined from a single spike train and does not rely on specifications of the input. Since the ISI irregularity is interesting on its own, I will refer to it as the *weak independent-spike hypothesis*: Even though neuronal spike generation is surely not a Poisson process for a given time course of synaptic inputs, the irregularity of discharge then implies that the correlation time of the intensity function has to be substantially smaller than the mean interspike interval. It is the weak independent-spike hypothesis, which I have taken as an argument to consider the relevant time scale to be about $1 - 10$ ms rather than $10 - 50$ ms. Note that the Bernoulli coding hypothesis came out mainly as a consequence of the *weak* independent-spike hypothesis, while it does not require the strong one to be true.

11.2 Faithful signal transmission

This section addresses the second point of the listing above and aims to clarify the functional motivation behind the Bernoulli coding hypothesis. I think, the main intuition underlying the idea of a rate code is not the statistical independence of spikes in the first place, but the *possibility of faithful signal transmission*. In contrast to the use of 'rate coding' as a strawman for a slow time scale, faithful signal transmission denotes the objective to faithfully represent the time course of the input signal. More precisely, it aims to achieve an output signal, which looks as similar as possible to the input signal in particular at the relevant scale of say $1 - 10$ ms.

In order to demonstrate that faithful signaling is not at all a demand that could be easily fulfilled, consider the PSTH response of a population of integrate-and-fire neurons that are driven by shot noise input currents which all have a step intensity function. As one can see in Fig. 11.2, the neuronal output does not faithfully reflect the input signal. Note that faithful signal transmission is a much stronger demand than that neurons can react rapidly. Since it was a wide-spread belief that the cut-off frequency of neuronal population rates is set by the membrane time constant of the neurons, it has been pointed out in the literature on population dynamics [GK02] that (integrate-and-fire) neurons react to changes in the input with a delay

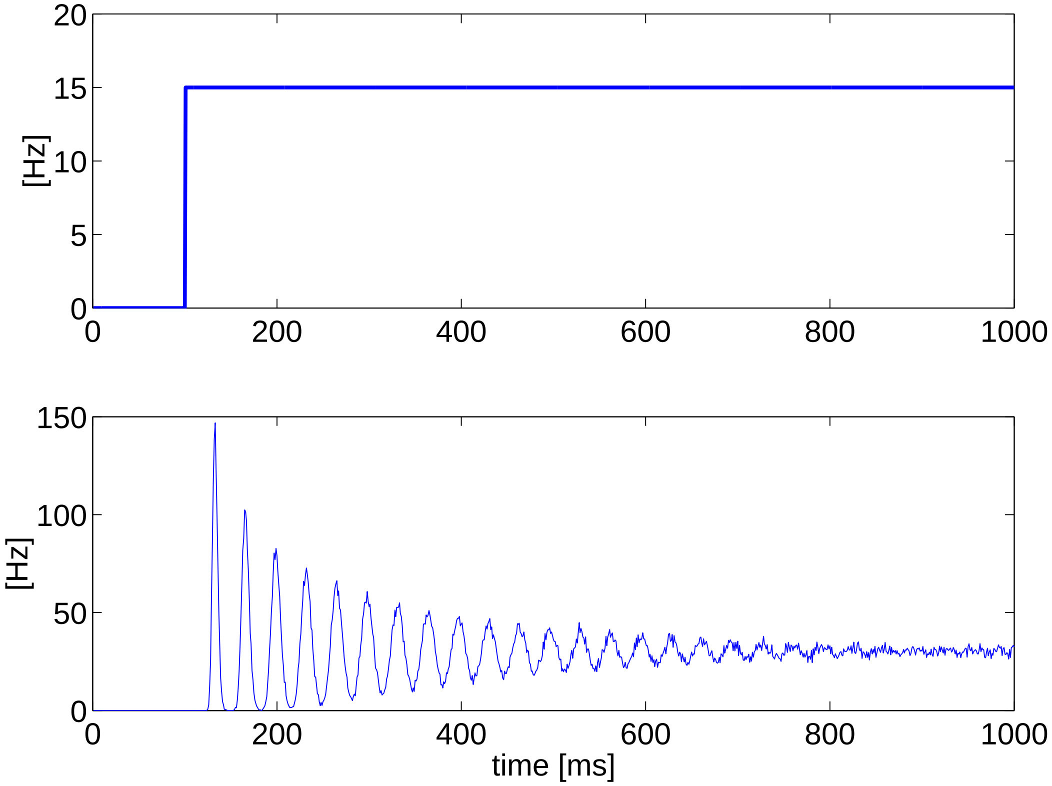

Figure 11.2. PSTH response of a population of integrate-and-fire neurons that are driven by shot noise input currents, which all have a step intensity function.

smaller than the membrane time constant. In Fig. 11.2, for instance, the rise time of the neuronal response after the step is way below the membrane time constant of the neurons. This, however, is not enough for faithful signal transmission!

In contrast to the population of integrate-and-fire neurons, large populations of Poisson neurons are capable to achieve faithful signal transmission. The temporal precision of the Poisson model is not limited in principle, but the time scale at which the intensity signal $\mu(t)$ within an arbitrary interval (t_1, t_2) can reliably be estimated decreases proportionally to the mean of the counting statistics $\lambda = \int_{t_1}^{t_2} \mu(t)dt$ without any limit. At the level of populations it is therefore well possible to achieve high temporal precision simply by a superposition of the individual spike trains. This is the main argument pointed out by various researchers why the Poisson process is compatible with high temporal precision [GK02; DA01].

11.2.1 Faithful signal transmission, computation, and temporal coding

At this point I should emphasize that I do not believe that neuronal functioning in the brain is about faithful signal transmission. The study of faithful signal transmission originates rather from a Gedanken experiment, which is motivated by the demand to have a self-consistent code that, *at a minimum*, should be able to 'compute' the simple function $f(x) = x$. In other words, even if neuronal processing clearly requires nonlinear computations, it appears reasonable to require that the realization of the identity function is not impossible.

As a theoretical argument the demand of faithful signal transmission is a strong hypothesis, on the basis of which one can rule out a lot of candidate coding schemes. Unfortunately, however, it is the property of being only a Gedanken experiment, which makes experimental testing so difficult: Of course, within neural networks in the brain, we clearly cannot expect to observe faithful signal transmission because of the network dynamics and its objective to *process* information rather than to 'copy' it.

The problem encountered in this example is very fundamental in neural coding and inevitably leads to severe difficulties in experimental testing. In fact, faithful signal transmission is quite the opposite of the definition of temporal encoding as has been proposed in [TM95]. As pointed out above, however, the *principle possibility* of faithful signal transmission cannot make a prediction about what it is likely to measure in experimental recordings. From a dynamical systems point of view, it is, in fact, almost impossible to avoid memory effects in a neural network. In this sense, temporal coding constitutes more or less a truism. The only reason, why it is possible that this is not obvious from experimental studies is due to the severe lack of determinism (i.e. the possibility of control) in the experiments, probing the neural code. I do not say this in order to criticize the experiments. Of course, if I knew how to do it better, I would just do it rather than talk about it. It is only important not to forget that the totality of data available to date includes a strong bias caused by the limited possibilities of experimental investigation. The latter might sometimes be obscured because of the large number of clever inventions and the variety of sophisticated techniques we already have.

I believe that the large variability is indeed due to our ignorance of the “real” functioning of the neural system, because under controlled input current conditions in slice experiments the reliability of spike timing is much higher [MS95], and complex dynamic systems are difficult to separate from contextual influences. The large gap between the principle signal transmission capabilities of neurons and the large amount of noise found in experiments supports the conjecture that, to a large extent, the variability may be explained by some other relevant variables that have simply not yet been discovered to date.

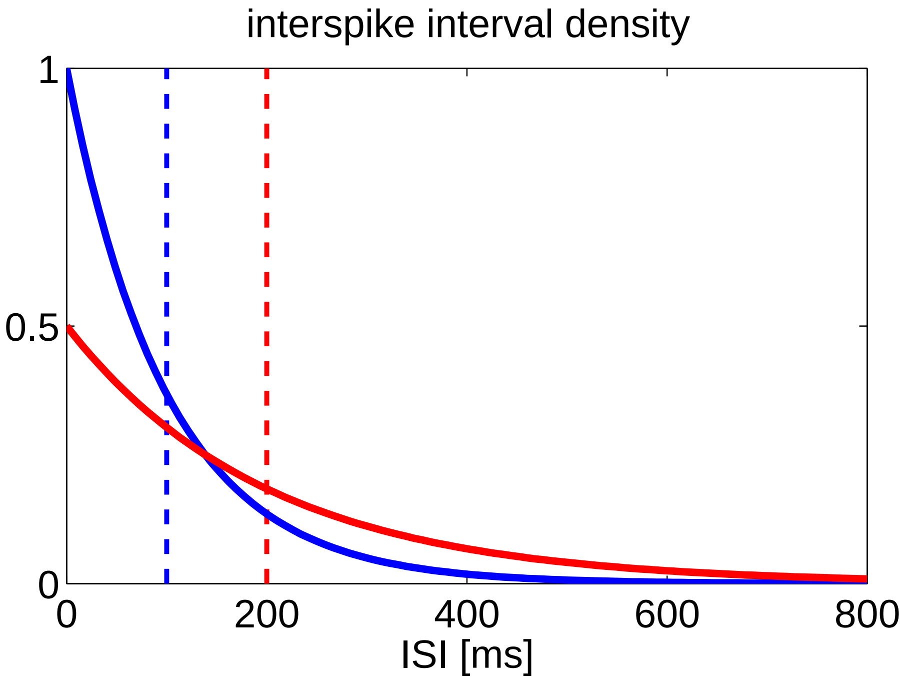

Figure 11.3. In case of a constant input, a deterministic neuron as the LIF or PIF neuron fires perfectly regularly, such that its ISI density is a delta function (dashed). Therefore, two different inputs (blue vs red) can be perfectly discriminated. The corresponding densities in case of a Poisson process are exponentially shaped (solid) and hence change only slightly in case of different inputs.

11.2.2 When information is noise

As we have seen in the example of population rate coding with Poisson neurons, the independent-spike hypothesis is different from the question of temporal precision. Nevertheless, the two issues are clearly not independent of each other. In fact, it appears that one has to sacrifice temporal precision for the independence of spikes, which can be illustrated by a comparison of the interspike interval distribution of the integrate-and-fire model with that of the Poisson model (Fig. 11.3): for constant inputs, the length of an interspike interval uniquely determines the strength of the input in case of the deterministic LIF or PIF model, but it is highly ambiguous when determining the input of a Poisson spike train. In other words, in case of the latter all irregularity in spike timing reflects nothing but noise, and it is necessary to average over many interspike intervals in order to achieve a reliable estimate of the input strength.

Now we want to compare the temporal precision of the Poisson model with that of the deterministic integrate-and-fire model in case of a time-dependent input signal. For the sake of clarity, we will consider the PIF model instead of the LIF model, which allows for a direct comparison with the Poisson model. As for any point process, the strongest constraint on information transmission in the PIF model stems from the translation of an analog input signal into a point process with a limited average firing rate: clearly, it is impossible to resolve an analog input signal at a time scale that is smaller than the corresponding interspike intervals. In contrast to the Poisson model, however, the PIF model is capable to get very close to this principle limit: each interspike interval T_k^{ISI} perfectly[3] reflects the average input within the last interspike interval (t_{k-1}, t_k):

$$\langle V \rangle_k = \frac{1}{t_k - t_{k-1}} \int_{t_{k-1}}^{t_k} V_{in}(t)\, dt = \frac{V_{th}\, \tau_{mem}}{T_k^{ISI}} \quad . \tag{11.5}$$

Therefore, the PIF neuron allows for a simple decoding of the input, which is minimax with respect to the mean squared error loss.

The Poisson model can be obtained as a stochastic version of the PIF model by introducing a particular sort of threshold noise: After each spike the threshold is randomly drawn from an exponential distribution with a mean that is equal to the fix threshold V_{th} of the noise free PIF model[4]. Consequently, in comparison with the PIF neuron, the representation of the input signal by a Poisson neuron is additionally impaired due to the variance of the threshold noise. A reliable estimate of the average intensity is only possible at a time scale that corresponds to a large number, say about a hundred, of interspike intervals.

In summary, the Poisson model appears to be quite disadvantageous for neuronal signal transmission. The comparison, however, did not take the limited average firing rate of real neurons into account. In the cortex, average rates of neurons are typically about 10 Hz or even lower. This implies that the average precision of a faithful code cannot go beyond 100 ms even for the perfect integrate-and-fire neuron. Since this is probably not precise enough, the two cases both, Poisson as well as the PIF model, need to make use of some sort of population coding in order to allow for a further reduction of the time scale. In case of Poisson neurons, it is equivalent to either multiply the intensity function of a single neuron by a number N in order to increase the temporal precision or to superimpose the spike trains of N neurons. This equivalence, however, does not hold true for the PIF neuron. Without any mechanism for decorrelating the activity of different neurons in the population, no additional information can be transmitted by more than one neuron. The average population response, in particular, then equals the shape of a single spike train and

[3] In case of the LIF neuron, the mapping from the average input within an interspike interval onto the length T^{ISI} of this interspike interval is not one-to-one anymore and the density $\rho(\langle V \rangle_k | T^{ISI})$ over the possible average inputs $\langle V \rangle_k$ depends on the waveform of the input. Precisely, the maximum possible degree of ambiguity can again be bounded by the inequalities displayed in Fig. 2.3.

[4] This choice ensures the average rates to be equal

does not realize a faithful signal transmission.

In conclusion, in population coding, neuronal noise can be useful in order to decorrelate the responses of different neurons. For faithful analog coding, in particular, the introduction of noise is *the only* strategy to make the spikes of different neurons sufficiently independent. In fact, I suspect it was the idea of faithful intensity coding, which lead Shadlen and Newsome [SN98] to the conjecture that the balance of excitation and inhibition may be required as a noise generator in order to achieve this goal. This model will be analyzed in more detail in the next section, before it will finally be demonstrated that label-pattern coding is not only more efficient but as a by-product can solve the problem of faithful signal transmission at the same time without the unattractive incorporation of noise.

11.3 Faithful intensity coding

Apart from signal-to-noise issues, population codes are fundamentally constrained by the neuronal dynamics. In particular, the biophysical properties of individual neurons as well as collective phenomena may substantially limit the speed at which a graded signal can be faithfully represented by the activity of an ensemble [Kni72]. Therefore, we now investigate as to how intensity encoding can be realized in populations of neurons, such that faithful signal transmission is possible. To this end, two rather general strategies are compared: depending on the ratio of excitatory input and inhibitory input, the input signal can be either encoded in the *population mean* or the *population variance* of the neuronal input currents.

In other words, in contrast to the usual characterization of a signal by the *temporal* mean and the variance components, we consider here the instantaneous distribution of input currents into the neurons of a functional ensemble at each moment in time. In this case, the synaptic inputs can also be divided into two components: one component is given by the input averaged over the ensemble; the other component is given by the deviations of individual inputs from the average: the *population variance*. Both components can, in general, fluctuate with time, contributing to the observable fluctuations in the synaptic currents of single neurons. The output of the ensemble is described by a PSTH. The population rate depends on the amplitudes of both components of the input. This reasoning allows one to conclude that signals delivered to a neuronal ensemble could in principle be carried by (encoded in) either the common, correlated, part of the synaptic inputs to the neurons, or the variance of the inputs across the population, or in both.

11.3.1 Theoretical analysis based on integrate-and-fire neurons

In order to assess the principle possibilities of these two strategies for faithful signal transmission, we consider an (infinitely) large population of identical integrate-and-fire neurons indexed by i, each receiving fluctuating synaptic inputs of the following form:

$$I_i(t) = \tilde{\mu}(t) + \tilde{I}_i(t). \tag{11.6}$$

Here $\tilde{\mu}(t) := \frac{1}{N}\sum_{k=1}^{N} I_k(t)$ stands for the instantaneous amplitude of the averaged, correlated part of the input (common to all the neurons), while $\tilde{I}_i(t) := I_i(t) - \tilde{\mu}(t)$ are the deviations of individual inputs from the average (unique for every neuron). A global measure for the time-dependent input diversity is given by the instantaneous *population variance* of the input $\tilde{\sigma}^2(t) = \frac{1}{N}\sum_{k=1}^{N} \tilde{I}_k^2(t)$. In general, both of the variables, $\tilde{\mu}(t)$ and $\tilde{\sigma}^2(t)$, change in time and thereby can serve as signals carrying information from presynaptic populations. The main goal of the analysis is to estimate how the activity of the population reflects the signals carried by either variable. We are, in particular, interested in the conditions under which the instantaneous population rate will faithfully follow the analog value of the signal *at any time.*

The analysis can be pursued to the ultimate solution, if $\tilde{I}_i(t)$ are mutually independent and temporally uncorrelated random processes (i.e. Gaussian white noise). In that limiting case $\tilde{\sigma}^2$ diverges, because of the vanishing correlation time. It is understood, however, that if the correlation time τ_c of a real current is sufficiently small (i.e. clearly smaller than the membrane time constant) essentially only the product $\sigma^2 = \tilde{\sigma}^2 \cdot \tau_c$ is relevant. For the sake of convenience, we will therefore call σ^2 somewhat incorrectly a "*population variance*" in the following as well.

We can thus write down the equation for the input currents

$$I_i(t) = \mu(t) + \sigma(t)\eta_i(t), \tag{11.7}$$

where $\mu(t)$ denotes the mean input across the population at time t, $\eta_i(t)$ is Gaussian white noise with unit spectral density, and $\sigma(t)$ is a scaling factor measuring how strongly the individual input currents deviate from the mean. Rigorously speaking, Gaussian white noise does not exist in nature, but it is understood as a reasonable idealization that together with the circuit equation (cf. Eq. 2.1)

$$\tau \frac{dV}{dt} = -(V - V_{rest}) + R_{in} I_{syn},$$

becomes a well defined mathematical term in the sense of Langevin equations.

Since the momentary state of an integrate-and-fire neuron is defined by only one variable – its membrane potential $V(t)$, the state of the whole population is completely characterized by a probability density function $P(V;t)$, i.e. the fraction of neurons with membrane potential close to V. The time evolution of the density function is governed by both the average and the fluctuating components of the inputs via the so-called Fokker-Planck equation [H.84]

$$\frac{\partial P(V,t)}{\partial t} = -\frac{\partial J(V,t)}{\partial V}\,, \tag{11.8}$$

where the probability flux $J(V,t)$ is given by

$$J(V,t) = \frac{R_{in}\mu(t) - V(t)}{\tau}\, P(V,t) - \frac{R_{in}^2\sigma(t)^2}{2\tau^2}\,\frac{\partial P(V,t)}{\partial V}\,. \tag{11.9}$$

The flux consists of two components: the first term is the drift component, which is governed by the mean $\mu(t)$, and the second term is the diffusion component, representing the effect of random fluctuations. The firing threshold and the reset potential impose boundary conditions on the flux, which imply $P(V_{th},t) \equiv 0$ for all $D = \frac{\sigma(t)^2}{2} > 0$ [TS95; H.84]. Therefore, at threshold, the only contribution to the flux is given by the diffusion component.

One can visualize this formulation by considering a collection of point-like particles moving independently along a one-dimensional axis under the combined influence of the deterministic force and a random, diffusing force, which tends to equilibrate the particles along the axis. In this analogy, $P(V)$ is just the density of particles on the axis of V.

The advantage of this approach is that one can derive an exact expression for the instantaneous firing rate of a population of neurons, which is given by the flux of particles to the firing threshold V_{th}:

$$R(t) = -\frac{R_{in}^2\sigma(t)^2}{2\tau^2}\,\frac{\partial P(V;t)}{\partial V}\bigg|_{V=V_{th}} \tag{11.10}$$

The instantaneous population rate $R(t)$ is here defined as the number of spikes emitted by the whole population in an (infinitely) small time bin around time t, normalized by the duration of the bin. Importantly, if the uncorrelated noise has a positive amplitude, only the diffusion component of the flux is contributing to the firing rate, since in this case, the density of particles at the threshold is zero. The above formula, together with the Fokker-Planck equation, represents a complete characterization of the population response in its dependence on the parameters of the input. In a stationary situation, when μ and σ^2 are constants, the density

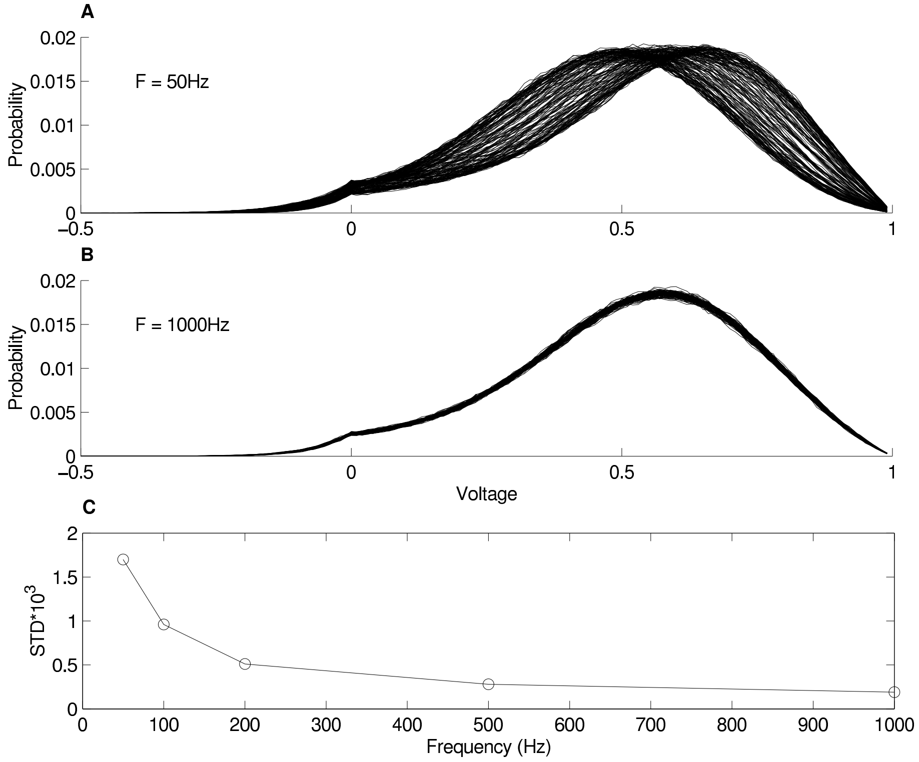

Figure 11.4. Simulations of the probability density function evolution in response to fast changes in input mean. A population of 20000 integrate and fire neurons with $\tau = 10$ ms, $V_{th} = 1$, $V_{rest} = V_{reset} = 0$ are simulated, with constant input variance $\sigma = 0.3$ and $\mu(t)$ oscillating between the values of 0.4 and 0.9, applied for 200 ms. (A) Overlayed probability density functions computed for subsequent 1 ms time bins, where $\mu(t)$ oscillates with a frequency of 50 Hz. (B) Same as A, but now $\mu(t)$ oscillates with a frequency of 1000 Hz. (C) The standard deviation of the density function averaged over positive values of voltage, plotted as a function of the frequency.

function adjusts its shape to the values of these parameters. The output rate, as stated above, is therefore a function of both parameters [RS69; Tuc88].

In the general situation, when the signals depend on time, the density function $P(V;t)$ evolves according to the Fokker-Planck partial differential equation, thereby exhibiting a low-pass filtered response to changes in the signals $\mu(t)$ and $\sigma^2(t)$ with a filter whose parameter depends on the membrane time constant of the neurons in a population [BH99]. This means that the output rate $R(t)$ given by Eq. (11.10) depends not only on the current values of the input parameters but also on their previous values i.e. the instantaneous population rate does *not* faithfully reflect the instantaneous signals. However, if the modulations of the input signals around some baseline values are much faster than the membrane time constant, the shape of $P(V;t)$ will be almost stationary, due to the above mentioned filtering property of the Fokker-Planck Equation.

To verify that the density function is indeed stable to very fast modulations of the input signals, we performed numerical simulations of the activity of a large population of integrate-and-fire neurons, receiving noisy input currents with an instantaneous mean value, oscillating in time with increasing frequency. The time evolution of the density function was then computed from the results of the simulations and its stability was assessed. In Fig. 11.4 A+B, we show a sample of probability density functions across the population computed over subsequent 1 ms bins for two different values of the frequency. Obviously, the variations in the density function are smaller for higher frequency. To quantify this result, we plot on Fig. 11.4 C the standard deviation of the density function, averaged over all positive values of the voltage, as a function of the frequency of the signal. For very high frequency, the standard deviation reduces to a low residual value that is explained by the finite size of the neuronal population. Qualitatively similar results were obtained for oscillating instantaneous variance in the input current (results not shown).

This result, together with Eq. (11.10), implies that the output population rate $R(t)$ will be proportional to the quickly changing instantaneous values of the population variance of the uncorrelated component of the inputs, $\sigma^2(t)$, and ignores the modulation in the mean. In other words, the population can faithfully transmit bandpass signals, if they are encoded in the amplitude of the uncorrelated noise.

Intuitively, this effect can be understood most easily in a population of neurons simultaneously receiving excitatory and inhibitory inputs. An excess of excitation at a given time will drive the neurons towards threshold. Some neurons will be caused to fire and immediately afterwards will be synchronously refractory. In order to achieve a rapid response, the excitatory pulse would have to be huge, which in turn would temporarily saturate the population activity. Therefore, the analog signal cannot be transmitted this way since the population activity will not be able to remain constantly at the intermediate values dictated by the signal. If, in contrast, the inhibition is increased simultaneously together with the excitation, this implies an increased population variance of the inputs. In this case, only a fraction of the neurons will receive big excitatory pulses and thereby respond rapidly, while a large proportion will be less affected or driven away from threshold. In this way, the balance of excitation and inhibition randomly selects changing subsets of neurons that are driven across threshold. This mechanism avoids population saturation and the number of firing neurons at each moment in time (population rate) faithfully reflects the graded signal encoded in the population variance of the input.

11.3.2 Experimental testing

In order to test the validity of this theoretical prediction, a series of experiments have been performed by Gilad Silberberg [SBM+04], using neocortical slice preparations. Experimental test is clearly necessary due to the many simplifications in the mathematical model, most notably the neglect of the kinetics of ionic channels. To this end, an ensemble of virtually white noise current traces (see caption Fig. 11.5) characterized by particular $\mu(t)$ and $\sigma^2(t)$ were injected into neocortical neurons while monitoring their spiking response. More precisely, it was repeatedly injected into single neurons, but every time choosing a *different* current trace from the same ensemble. The 'population' activity has then been estimated by using a PSTH with time bins of 1 ms.

In the first series of experiments, the response to two simple forms of input signals are compared for which either $\mu(t)$ or $\sigma^2(t)$ increases abruptly at a particular time (Fig. 11.5 A+B). Eq. (11.10) predicts that there should be a gradual response to the jump in μ, and an instantaneous initial response to a change in σ^2 since the latter enters as a multiplying factor in the expression for the response. The amplitudes of the signals were calibrated such that the steady-state level of a neuron's firing was identical in both cases, because this implies that the average firing rates are almost identical in both cases. However, the time profile of the PSTH in response to abrupt increase in μ and σ^2 was very different: while in the first case, there was a gradual change in the PSTH, in the second case we observed an instantaneous initial response with subsequent decay to the new stationary level. These results are in agreement with the theoretical prediction (Eq. (11.10)).

Similar results were obtained in all 18 of a wide variety of neocortical neurons, which included pyramidal neurons and different types of interneurons. It has to be emphasized that in both cases, the output rate continued to change after the transition, when the input signal (the values of either μ or σ^2) stayed constant at a different level. This means that the instantaneous population rate did not follow the step in the signal amplitude when this was modulating the mean current to the population. Therefore, in both cases, the responses did *not* faithfully reflect the overall shape of the single step and showed temporal modulations, which are also a prediction of the theory presented above.

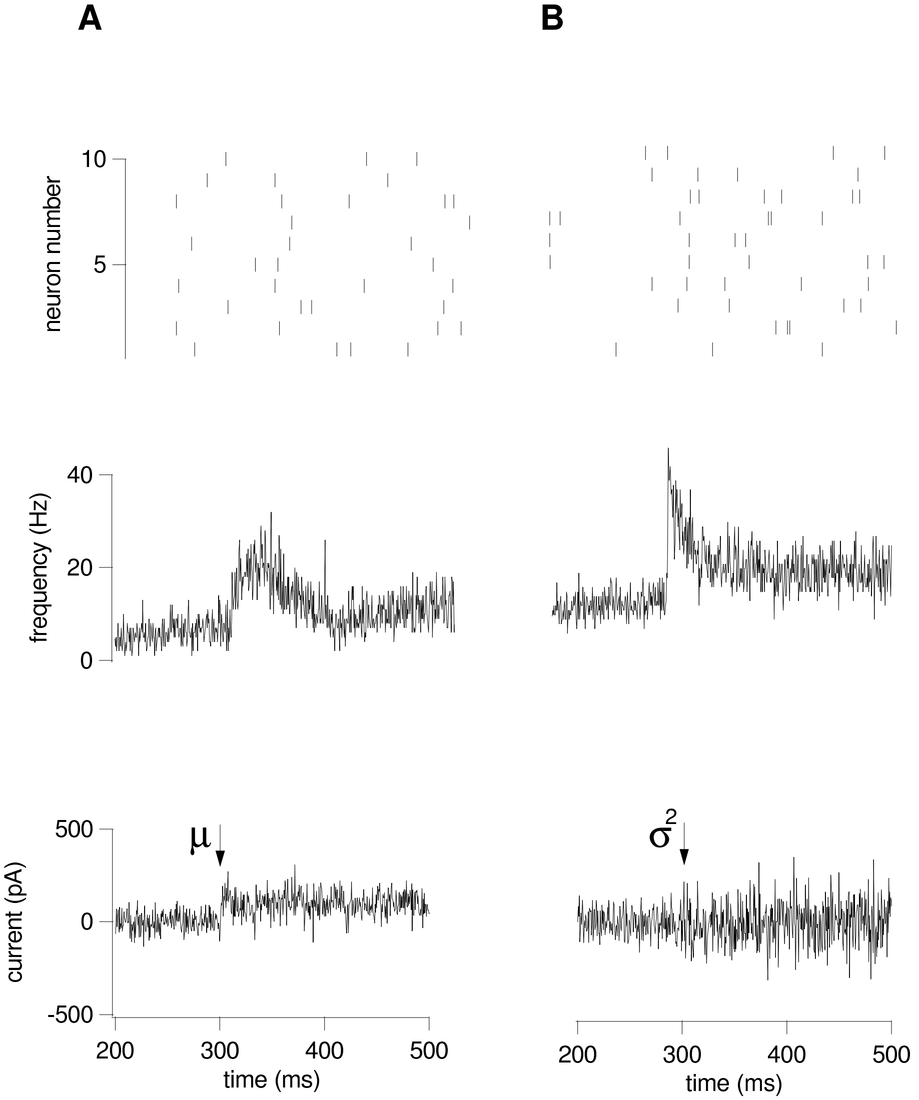

Figure 11.5. Response of neocortical neurons to abrupt changes in input parameters μ and σ^2. In each case, 4000 different virtually white noise current traces (sampling interval in all of the experiments is $\tau_{sample} = 0.25$ ms) were injected into a pyramidal neuron sequentially. (A) Lower trace, an example of a single current trace injected. The arrow indicates the moment at which the amplitude of the mean current was increased. Middle trace, histogram of the "population"response with a time bin of 1 ms. At the transition point, the mean current was increased from 120 pA to 200 pA. Upper trace - raster plot of spike trains for 10 randomly chosen trials. Solid red lines show the stationary levels of the response before and after the transition. (B) Same as in A, but the set of current time-courses featured a change in the population variance $\sigma^2 = \tilde{\sigma}^2\tau_{sample}$ from 22.5 (pA)^{2}sec to 90 (pA)^{2}sec. Histogram binning is 1 ms.

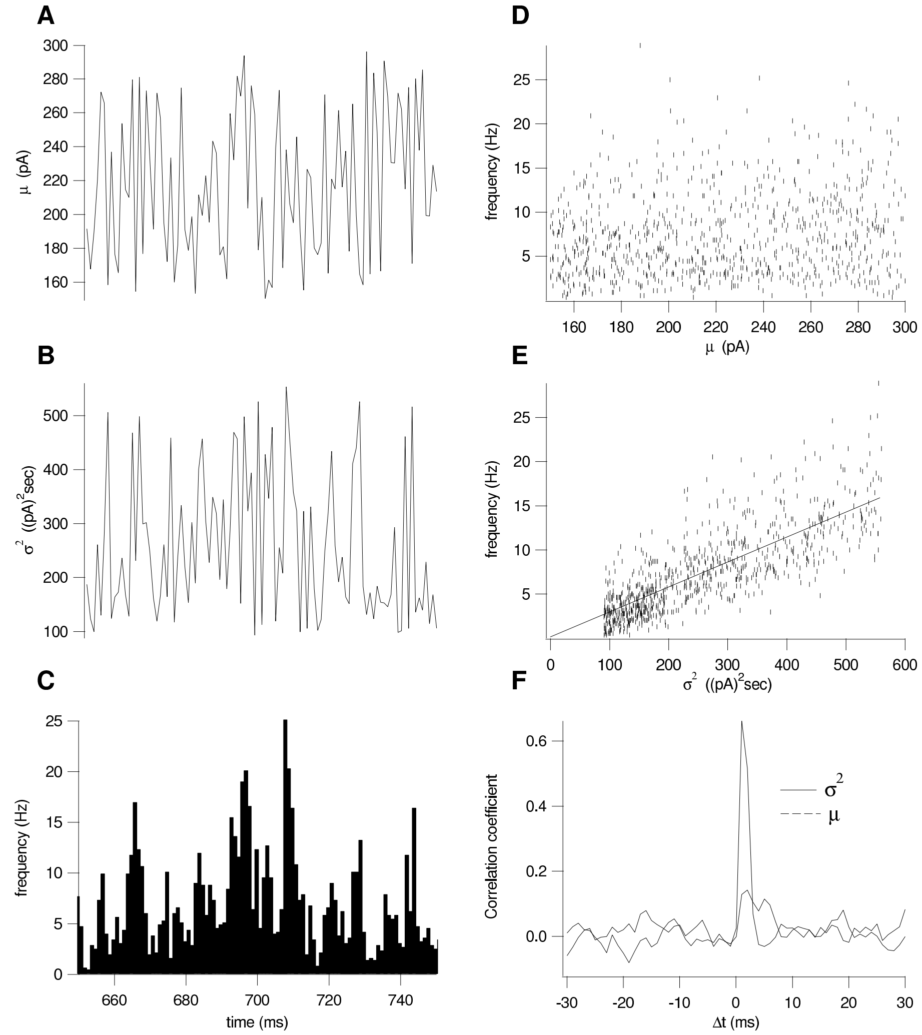

Figure 11.6. Response of a pyramidal neuron to virtually white noise current injection (τ_{sample} = 0.25 ms), in which two temporally uncorrelated signals are encoded. (A) Instantaneous values of mean current. (B) Instantaneous values of population variance. (C) Instantaneous response (PSTH) in time bins of 1 ms. For every time bin, the PSTH was computed from the spiking responses of a neuron to repeated injections of different current traces as explained in the text. (D+E) Instantaneous response plotted vs the instantaneous value of μ and σ^2, respectively, after compensating for a 1 ms delay. Every point on the graph represents a pair ($\mu(t)$, PSTH(t)) or ($\sigma^2(t)$, PSTH(t)) for consecutive time bins. (D) Correlation coefficient R = 0.15. (E) Red line is a linear regression. The sample correlation coefficient is $\rho = 0.79$ and the sample correlation ratio is 0.65, which is very close to ρ^2 and hence confirms the linear relationship [SO94]. Overall, 5600 current traces were injected (i.e. the observed correlation reflects the precision that would be achieved in a population at a time scale of 1 ms in a population of about 5000 neurons. (F) The cross-correlation values between the "population"response and the signals carried by the mean and the variance of the input currents.

The main theoretical prediction then has been tested by studying the population response to signals that undergo rapid *ongoing* changes in time. To this end, the same type of current forms have been injected as in the first experiment, while now both, $\mu(t)$ and $\sigma^2(t)$ are rapidly fluctuating all the time. In other words, the two signals were present simultaneously in the input. In order to separate the effectiveness of these two signals, we constructed $\mu(t)$ and $\sigma^2(t)$ by randomly assigning new values to each of them independently at every millisecond drawn from uniform distributions (see Fig.11.6 A+B). The ranges for both kinds of input signals were calibrated for each neuron individually, such that the corresponding output ranges of the firing rates were of the same size (i.e. the calibration makes sure that the stationary firing rate for constant μ, σ^2 is identical in both cases, $(\mu = \mu_{min}, \sigma^2 = \sigma^2_{max})$ and $(\mu = \mu_{max}, \sigma^2 = \sigma^2_{min})$.

The PSTH response to these currents was obtained (Fig. 11.6 C). Strikingly, the instantaneous population rate reliably followed the signal carried by the variance of the input currents, and there was no observable correlation with the signal contained in the mean (Fig. 11.6 F). To quantitatively test the prediction contained in Eq. (11.10) about the dependency of the response on the components of the input, we plotted the instantaneous values of the PSTH response vs the instantaneous variance of the currents (Fig. 11.6 E). In agreement with the prediction, the points in this graph scatter around the straight line passing through the origin. The same graph, but with the signal contained in the average current μ, does not result in any significant dependency (Fig. 11.6 D), again following the theoretical prediction. Very similar results were obtained in all 4 pyramidal neurons, which were tested.

The above experiments confirm the validity of the abstract mathematical analysis. However, because in these experiments the injected currents were artificial and, in particular, did not correctly reflect the temporal correlations of real input currents, he, in addition, obtained realistic synaptic currents from whole-cell voltage-clamp recordings in cortical slices with different levels of excitation. With these currents, a control experiment has been performed by injecting them into a neuron and recording the discharge responses for both, step changes in mean current as well as step changes of the variance.

Like in the first experiments, the discharge response increased gradually when the recorded currents were injected, containing abrupt changes in the mean current (Fig. 11.7 A). In contrast, the change in discharge was much faster when the variance of the current increased abruptly (Fig. 11.7 B). These results demonstrate that the difference between signaling by variance and signaling by mean persists with real synaptic noise currents (c.f. [BCFA01]). In particular, signaling by variance does not require white noise currents, but is possible with the synaptic currents generated in neocortex.

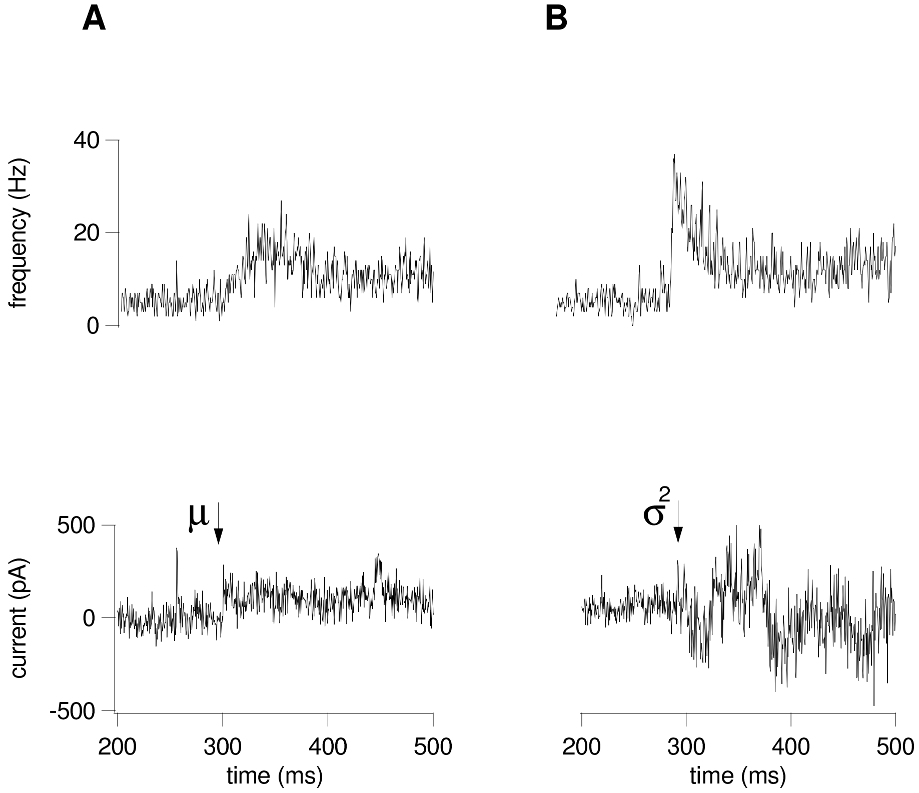

Figure 11.7. Response of neocortical neurons to currents obtained from voltage-clamp experiments. (A) Lower trace, an example of a segment of a current obtained in a voltage-clamp recording from a slice with low activity level. At the transition time, a constant value was added to the current. Upper trace, histogram of 4000 responses to different current segments. (B) Same as in A, but with a step change in the variance of the current. This was achieved by switching to current traces, recorded at higher activity level.

Conclusion

The analysis as well as the experimental tests have shown that faithful analog signal transmission is not easy to achieve in populations of cortical neurons. Eq. 11.10, in particular, determines conditions for which faithful signal transmission is possible. This is the case if the population variance of the input currents carries the signal to be represented and the latter, in addition, is a rapidly changing bandpass signal.

This finding matches the intuition of Shadlen and Newsome [SN98] that the balance of excitatory and inhibitory synaptic inputs is important for intensity coding. While such a balance can explain the high degree of irregularity in cortical spike trains [GM64], it does not support the conjecture that current fluctuations are nothing but noise, apart from their amplitude. From a functional point of view, in particular,

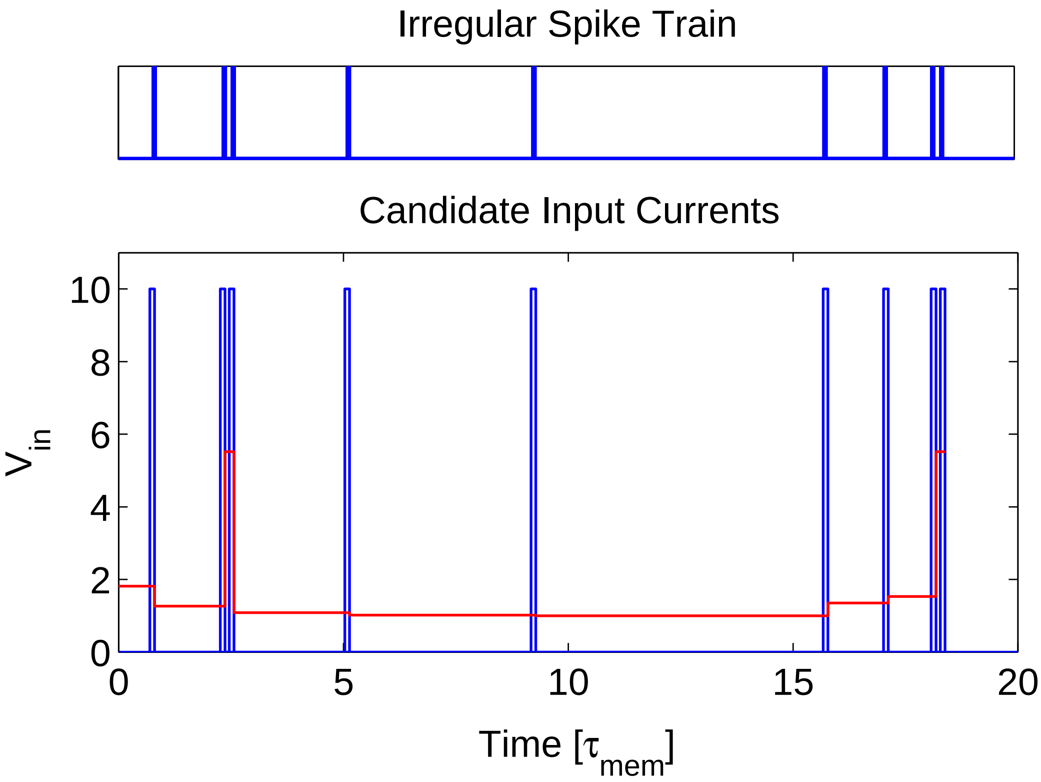

Figure 11.8. Ambiguity of the candidate input currents for a given spike train (upper). The lower plot shows two examples of input waveforms, which both may have generated the spike train, when fed into an integrate-and-fire neuron. While Bernoulli coding suggests to use waveforms, which are either at maximum or zero (blue), the other extreme (red) is a waveform, which is constant within each interspike interval, and hence minimizes the amplitude of the input current.

this idea appears to be not very attractive, because the required noise will clearly impair the coding efficiency, and the uncoordinated cancellation of action potentials constitutes a large waste of energy consumption. The last section will show that the superiority of Bernoulli coding over intensity coding is indeed even more obvious when taking the constraints of the neuronal dynamics into account.

11.4 Faithful Bernoulli coding vs intensity coding

From the (static) population coding point of view, the 'signaling by variance' strategy presented in the previous section is a randomized version of the encoding strategy presented in section 10.1.2 above. While there, the partitioning of the neurons into the two subpopulations of maximally activated neurons S_1 and silent neurons S_0 was a deterministic function of x, it is random in case of signaling by variance. In

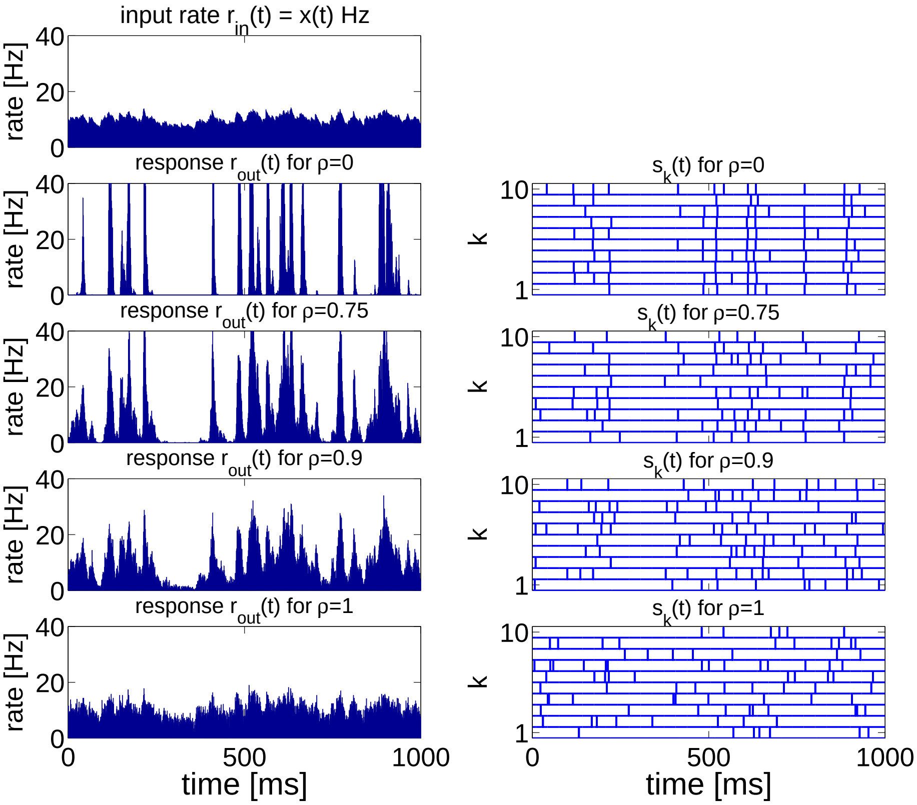

Figure 11.9. Balance of excitation and inhibition is needed for faithful signaling. With increasing balance ratio, the output signal (lower left) looks more and more like the input signal (upper left). For illustration, a sample of spike patterns is shown for each case ($\rho = 0, 0.75, 0.9, 1$) as well (right).

other words, signaling by variance is actually not a pure intensity code anymore, but it already resorts to a binary encoding strategy.

There is, in fact, a simple explanation, why binary encoding strategies are so important for faithful signal transmission: Since the absence of spike generation is not constraint by the history[5], the essential constraint on the possibility of faithful signal transmission stems from the time required to integrate over the synaptic inputs in order to generate an action potential. Because this time takes a minimum, if the input is maximal, it is an optimal strategy to use maximum current input for spike generation.

This insight allows us to restate the Bernoulli coding hypothesis, eliminating the reliance on the discrete-time memoryless channel. The crucial point is that any spike train can be generated in infinitely many ways. For the sake of clarity, consider the

[5] It is well justified to assume that a neuron does not fire if its input is zero, independent of its input in the past.

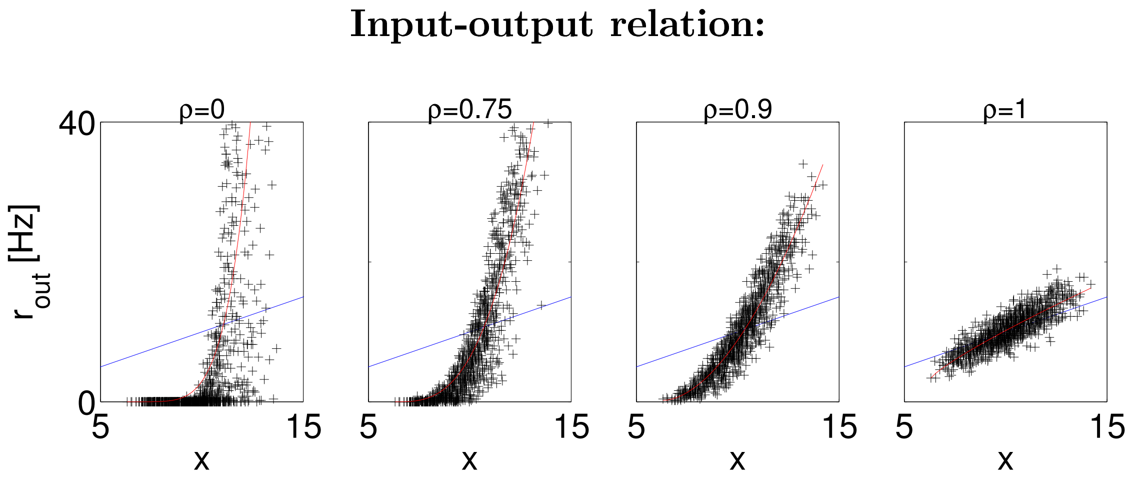

Figure 11.10. Input-output relationship for different degrees of balance. Crosses indicate samples from $P(r_{out}|r_{in})$. Red line corresponds to a tuning function and is obtained by algebraic regression. Blue line indicates the intended function $f(x) = x$. The minimal reconstruction error of $2.4 \cdot 10^{-1}$ is achieved for a balance ratio of $\rho = 0.9$.

example given in Fig. 11.8. The *Bernoulli coding hypothesis* is now completed by an additional statement of how to resolve the ambiguity of candidate input: it suggests that the *spike trains should be generated by current waveforms, where the individual spikes are caused by preferably short and hence as large as possible current pulses.*

In order to illustrate, the superiority of Bernoulli coding over intensity coding, a concrete example with $N = 5000$ integrate-and-fire neurons is discussed in the following. The task is to faithfully transmit a colored noise signal $x(t)$ with mean $\langle x \rangle = 10$ and a correlation time equivalent to the membrane time constant (say 10 ms for illustration) of the integrate-and-fire neurons. The firing threshold of the latter is chosen such that self-consistency for the average rates $\langle r_{in} \rangle = \langle r_{out} \rangle = 10$ Hz is achieved. The output signal $r_{out}(t)$ is given by the sample mean PSTH $r_{out}(t) = \frac{1}{N}\sum_{k=1}^{N} s_k(t)$, where $s_k(t)$ is the PSTH (bin=1ms) of neuron k.

In case of population rate coding, the signal x(t) is encoded in the instantaneous firing rate $r_{in}(t) := x(t) \cdot 1$ Hz of the presynaptic neurons that are modeled by an inhomogeneous Poisson process. There are N_e excitatory neurons and N_i inhibitory neurons in the presynaptic population, which in total $N_e + N_i$ equal the total number of postsynaptic neurons $N = 5000$. Four different degrees of balance are considered, characterized by the balance ratio $\rho = \frac{N_i}{N_e}$ (see Fig. 11.9) The case of $\rho = 1$, in particular, matches the coding strategy discussed at length in [SN98].

The performance of the encoding with respect to faithful signal transmission can be assessed by the instantaneous response plotted as a function of x (Fig. 11.10). The

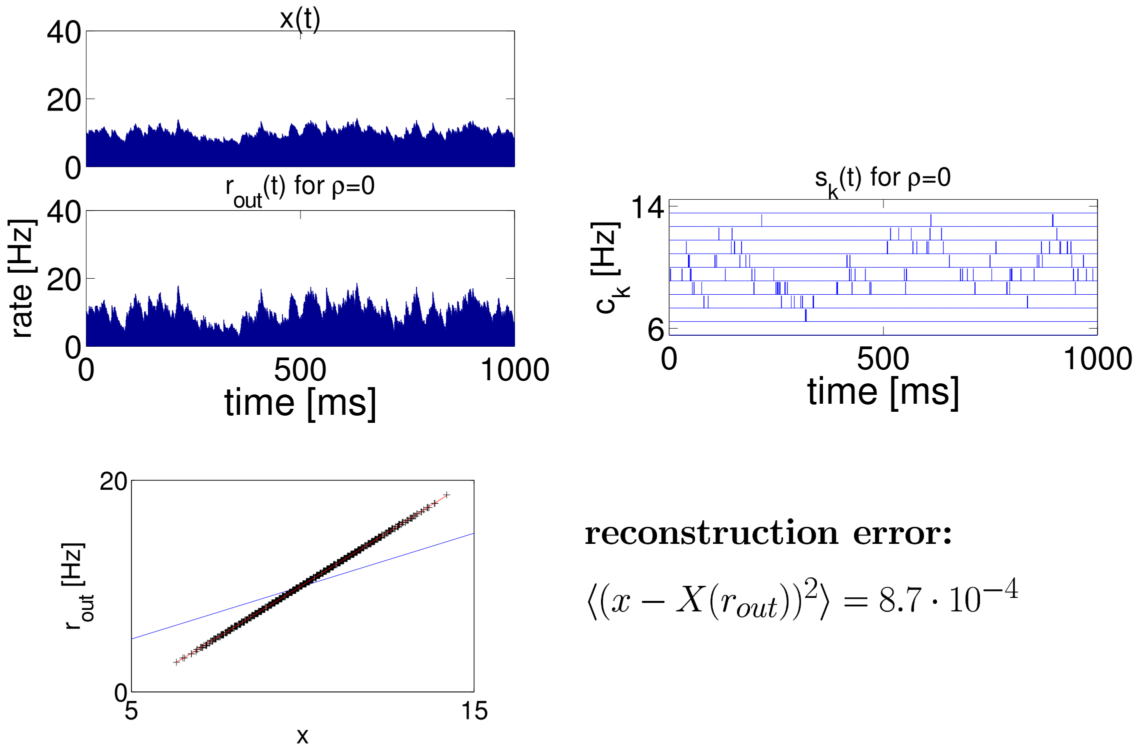

Figure 11.11. Binary encoding - analog decoding. The upper left shows a comparison of the input signal (upper) with the population average output signal. A sample of spike trains for neurons with different tuning function centers is displayed at the right. Furthermore, the input-output relationship is shown (lower left).

regressions (red lines) can be interpreted as the gain functions of the neurons: The larger the balance ratio, the closer these gain functions are to the intended function $f(x) = x$. The minimal reconstruction error of $2.4 \cdot 10^{-1}$ is achieved, however, for the balance ratio of $\rho = 0.9$.

As explained above, 'signaling by variance' is a randomized version of a binary encoding strategy. Clearly, we can expect a much better signal-to-noise ratio, if we use the deterministic version instead. This can be achieved, for instance, in the following way. Suppose each neuron k receives excitatory input from a box-like receptive field $r_k^E(t) = r_{max}\Theta(x(t) - c_k - 0.05)\,\Theta(c_k + 0.05 - x(t))$, where c_k denotes the center of the tuning function. Further assume that the centers are distributed according to a triangular distribution such that more neurons are activated in case of large x and less are activated in case of small x. Using this strategy, one obtains the result shown in Fig. 11.11. The minimal reconstruction error is more than three orders of magnitude smaller than in case of optimal balance. A further reduction of the minimal reconstruction error can be achieved, if the subsequent neuronal readout distinguishes between the labels of the different neurons. By using the simplistic population vector method and a uniform distribution of the tuning function centers the error has been reduced to $3.2 \cdot 10^{-7}$, which is about six orders of magnitude smaller than in case of optimal balance.

In conclusion, the examples considered here confirm the argument that Bernoulli coding is not only optimal with respect to 'spatial' decorrelation between different neurons, but is advantageous with respect to the temporal decorrelation of the individual neuronal responses as well.

Chapter 12

Conclusion

The objective of the quest for the neural code is the missing link between electrophysiology and function. Consequently, the search for the neural code allows for two complementary approaches, starting either from electrophysiological data or from a functional point of view.

In this thesis, the latter approach was taken, combining two areas of research – efficient coding and population coding – that have developed rather independently from each other. Conceptually, this study is motivated by the insights of efficient coding research, which emphasizes the importance of knowing the shape of the signal for the understanding of neuronal representations. The models investigated, however, have been mainly informed by biophysical constraints on neuronal signal transmission, as it is more common in the research on population coding. Particular emphasis is placed on the fact that the shape of efficient representations may critically depend on a multitude of constraints rather than to focus too much on a single principle. The somewhat lighthearted play with the ingredients of a model served the goal of this thesis, which was to achieve more educated intuitions and hypotheses about the neural code.

Starting from the standard model of population coding for the study of optimal tuning widths, diverging conclusions in the literature have been resolved by the introduction of a new independent parameter, namely the dynamic range of a tuning function (chapter 4). The difficulties of applying this standard model to neuronal representations of, say natural images, motivated a more exhaustive search for characteristic features of population codes that are most relevant for coding efficiency. Using Fisher information, the commonly used measure of coding efficiency in the context of population codes, the maximum reduction of the dynamic ranges of the tuning functions turned out to be most crucial for high efficiency (chapter 5, [BRP02]). At the same time, however, this less restricted optimization uncovered severe limitations of Fisher information as a measure for coding efficiency.

Several examples have been presented that help to assess which dependencies of Fisher information on the shape of a population code are meaningful and which are not. This experience suggested that a small dynamic range does not only lead to large Fisher information but is an important aspect of population coding also beyond the scope of Fisher information.

In order to not rely on the heuristic argument of Fisher-optimality, we then proceeded to test the advantage of label-pattern coding by explicit numerical evaluations of the minimum mean square error (MMSE) for a selection of characteristic examples (chapter 6). This study has not only confirmed the superiority of label-pattern coding above intensity coding, but it also constitutes the first 'exact' investigation of the minimum mean square accuracy of population codes in the literature [BRP03b].

Another advantage of using the MMSE in place of Fisher information is its clear correspondence to an explicit estimator, which can be derived from first principles. Instead of relying on the hegemony of Fisher information, it thus becomes possible to investigate how specific constraints on the decoding affect optimal encoding strategies. This possibility has been used in chapter 7 to test whether the advantage of label-pattern coding holds true also in case of a limited computational power of the neuronal readout. While strict linearity turns out to suspend the advantage of label-pattern coding, it is reinstated again if one allows for a simple threshold nonlinearity. Since the possibility of a threshold operation appears to be more than plausible in case of neurons, I concluded that strict linearity is likely to be too restrictive. This fact should also be kept in mind as a possible pitfall, when dealing with linear models for the sake of simplicity.

Alluding to the linear-nonlinear cascade neuron models, it is argued in chapter 8 that the optimization of the nonlinear gain functions appears to be the most unbiased approach, in order to deal with the situation of lacking knowledge about the particular function of a neuron. Again, this study confirms the advantage of binary coding at physiologically plausible time scales. While not so much important for the conclusion on the relevance of binary coding, the analytically proven existence of a phase transition from analog towards binary coding at a critical maximum mean spike count of slightly less than three is an intriguing piece of physics discovered on the side [BRP03c].

The totality of results on optimal population coding investigated in the first part of the thesis, led me to the proposal of the *Bernoulli coding hypothesis* (cf.[BRP03a]). In short, it states that analog coding does not suit rate coding at physiologically plausible time scales because of its distinct inferiority with respect to coding accuracy. In the second part, the efficiency argument took a back seat, but other criteria have been considered to challenge the Bernoulli coding hypothesis.

In chapter 10 the intuition is addressed that analog encoding strategies, like population rate coding, are more robust against cell death and other sources of unreliability.

This intuition is invalidated by a comparison, which shows substantial superiority of label-pattern coding, independent of the amount of noise and of the number of neurons randomly erased.

In chapter 11, finally, the discrete-time, memoryless neuron models are replaced by integrate-and-fire neurons in order to investigate the effect of the neuronal dynamics on the *possibility of faithful signal transmission* in case of analog and binary encoding strategies. The meaning of ’faithful signal transmission’ is discussed in detail, and, as I hope, will contribute some clarifying hints to the ongoing debate on ’rate coding’ vs ’temporal coding’.

The results of chapter 11 show that faithful signal transmission is, generally speaking, difficult in case of analog coding. Analog encoding necessarily implies a much higher degree of homogeneity in the neuronal population, so that it is necessary to introduce noise in order to decorrelate neuronal responses temporally and spatially. Inspired by the conjecture that the random-walk model of Gerstein and Mandelbrot [GM64] is well-suited for the realization of an analog population rate code with a large dynamic range [SN98], we investigated the possibilities of faithful signal transmission with respect to the two relevant input signal components in this model: the average population input (the *mean current*) and the *population variance*.

While signaling by the mean current does not allow for faithful signal transmission in general, it has been shown that signaling by variance may work for this purpose in case of rapidly fluctuating bandpass signals. Essential predictions of the abstract analysis have been confirmed in a series of experiments, using neocortical slice preparations.

A strikingly superior performance in faithful signal transmission is easily achieved, however, by using a label-pattern encoding strategy. Although this investigation has been carried out not in the same detail as the work on coding efficiency so far, the dynamical constraints obviously provide a strong argument for the Bernoulli hypothesis, maybe even stronger than that of coding efficiency.

Appendix A

Appendix

A.1 Exponential family

An s-parametric family of distributions is a member of the exponential family, if it can be written in the following form:

$$P(x|\theta) = P_0(x) + \exp\left\{\sum_{i=1}^{s} \eta_i(\theta)T_i(x) - B(\theta)\right\} \tag{A.1}$$

These distributions can always be written in the canonical form:

$$P(x|\eta) = P_0(x) + \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right\} \tag{A.2}$$

where $\eta = (\eta_1, \ldots, \eta_s)$ is called the natural parameter. Prominent examples of the exponential family are the normal family, the Poisson family, and the Multinomial family. The Multinomial family $M(\eta_1, \ldots, \eta_s; n)$ is given by

$$P(k|\eta_1, \ldots, \eta_s, n) = \frac{n!}{k_1! \cdots k_{s+1}!} \exp\left\{\sum_{i=1}^{s} k_i\eta_i - n\log\left(1 + \sum_{i=1}^{s} e^{\eta_i}\right)\right\} \tag{A.3}$$

where the natural parameters $\eta_1, \ldots, \eta_s \in \mathbb{R}$. Using $p_i := p_{s+1}e^{\eta_i}, i = 1, \ldots, s$ and $p_{s+1} := \left(1 + \sum_{i=1}^{s} e^{\eta_i}\right)^{-1}$ Eq. A.3 can be transformed into the more familiar form

$$P(k|p_1, \ldots, p_{s+1}, n) = \frac{n!}{k_1! \cdots k_{s+1}!} p_1^{k_1} \cdots p_{s+1}^{k_{s+1}} \tag{A.4}$$

If $s = 1$ one obtains the Binomial family and in the special case of $n = 1$ this gives the Bernoulli family.

A good thing about the members of the exponential family is that it is easy to obtain the moments and cumulants of their distributions, because the moment generating function is always given by

$$M_T(u) = \exp\left\{A(\eta + u) - A(\eta)\right\} \quad . \tag{A.5}$$

A.2 Monte-Carlo integration

In general, the evaluation of Eq. 5.6 or Eq. 6.5 requires integration over an N-dimensional space. According to the Monte-Carlo technique [Bis95], we estimate the value of Eq. 5.6 or Eq. 6.5 by an average over n trials, for which a particular $(x, \mathbf{k})_i$ is randomly drawn from the joint distribution $p(\mathbf{k}|x)p(x)$ with $p(x) = 1$ for all $x \in [0, 1]$ and otherwise zero. Then it holds:

$$\mathrm{E}[(\hat{x}(\mathbf{k}) - x)^2] \approx \frac{1}{n}\sum_{i=1}^{n}(\hat{x}(\mathbf{k}_i) - x_i)^2\,. \tag{A.6}$$

The error of this approximation decreases with the number of trials. We evaluated the r.h. side of Eq. A.6 up to the second relevant digit. As a termination criterion, the averaging process has been stopped , when there was no change in the value of the second relevant digit, during the last 10000 trials.

Bibliography

[AD99] L.F. Abbott and P. Dayan. The effect of correlated variability on the accuracy of a population code. *Neural Comput.*, 11:91–101, 1999.

[AH02] D.L. Adams and J.C. Horton. Shadows cast by retinal blood vessels mapped in primary visual cortex. *Science*, 298:572–576, 2002.

[AN00] S. Amari and H. Nagaoka. *Methods of Information Geometry.* Oxford University Press, New York, 2000.

[AS42] A.C. Aitken and H. Silverstone. On the estimation of statistical parameters. *Proc. Roy. Soc. Edin. A*, 62:369, 1942.

[ASGA96] A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science*, 273(5283):1812, 1996.

[Ati92] J.J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251, 1992.

[Att54] F. Attneave. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[BA91] V. Braitenberg and Schüz A. *The anatomy of the cortex. Statistics and Geometry.* Springer, Berlin, Heidelberg, New York, 1991.

[Bad96] R. Baddeley. An effeicient code in v1. *Nature*, 381:560–561, 1996.

[Bar59] H.B. Barlow. Sensory mechanisms, the reduction of redundancy, and intelligence. In *The Mechanisation of Thought Processes*, pages 535–539, London: Her Majesty's Stationery Office, 1959.

[BBdRvS00] N. Brenner, W. Bialek, and R.R. de Ruyter van Steveninck. Adaptive rescaling maximizes information transmission. *Neuron*, 26:695–702, 2000.

[BBH$^+$01] J. Benda, M. Bethge, M. Henning, K. Pawelzik, and A.V.M. Herz. Spike-frequency adaptation: Phenomenological model and experimental tests. *Neurocomput.*, 38-40:105–110, 2001.

[BC02] P.C. Bressloff and J.D. Cowan. Spontaneous pattern formation in primary visual cortex. In S.J. Hogan, A. Champneys, and B. Krauskopf, editors, *Nonlinear dynamics: where do we go from here?*, chapter 11. Institute of Physics, Bristol, 2002.

[BCFA01] N. Brunel, F.S. Chance, N. Fourcaud, and L.F. Abbott. Effects of synaptic noise and filtering on the frequency response of spiking neurons. *Phys.Rev.Lett.*, 86, 2001.

[Ber79] J.M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7:686–690, 1979.

[BH88] P. Baldi and W. Heiligenberg. How sensory maps could enhance resolution through ordered arrangements of broadly tuned receivers. *Biol. Cybern.*, 59:313–318, 1988.

[BH99] N. Brunel and V. Hakim. Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comput.*, 11, 1999.

[Bis95] C.M. Bishop. *Neural networks for pattern recognition.* Oxford University Press, New York, 1995.

[Bla88] R.E. Blahut. *Principles and practice of information theory.* Addison-Wesley, Cambridge, MA, 1988.

[BN98] N. Brunel and J.-P. Nadal. Mutual information, fisher information, and population coding. *Neural Comput.*, 10:1731–1757, 1998.

[BP01] M. Bethge and Pawelzik. Synchonous inhibition as a mechanism for unbiased selective gain control. *Neurocomput.*, 38-40:483–488, 2001.

[BPG99] M. Bethge, K.R. Pawelzik, and T. Geisel. Brief pauses as signals for depressing synapses. *Neurocomput.*, 26-27:1–7, 1999.

[BRP02] M. Bethge, D. Rotermund, and K. Pawelzik. Optimal short-term population coding: when fisher information fails. *Neural Comput.*, 14:2317–2351, 2002.

[BRP03a] M. Bethge, D. Rotermund, and K. Pawelzik. Binary tuning is optimal for neural rate coding with high temporal resolution. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, Cambridge, MA, 2003. MIT Press.

[BRP03b] M. Bethge, D. Rotermund, and K. Pawelzik. Optimal neural rate coding leads to bimodal firing rate distributions. *Network: Comput. Neural Syst.*, 14:303–319, 2003.

[BRP03c] M. Bethge, D. Rotermund, and K. Pawelzik. A second order phase transition in neural rate coding: binary encoding is optimal for rapid signal transmission. *Physical Review Letters*, 90:088104, 2003.

[Bru34] E. Brunswik. *Wahrnehmung und Gegenstandswelt: Grundlegung einer Psychologie vom Gegenstand her.* Deuticke, Leipzig, 1934.

[Bru43] E. Brunswik. Organismic achievement and environmental probability. *Psychological Review*, 50:255–272, 1943.

[BS97] A.J. Bell and T.J. Sejnowski. The ”independent components”of natural scenes are edge filters. *Vision Res.*, 37(23):3327–38, 1997.

[BSK$^+$00] N. Brenner, S.P. Strong, R. Koberle, W. Bialek, and R.R. de Ruyter van Steveninck. Synergy in a neural code. *Neural Comput.*, 12:1531–1552, 2000.

[Chi01] E.J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Comput. Neural Syst.*, 12(2):199–213, 2001.

[Cra46] H. Cramér. A contribution to the theory of statistical estimation. *Aktuariestidskrift*, 29:458–463, 1946.

[Cra99] M.C. Crair. Neuronal activity during development: permissive or instructive? *Curr. Op. Neurobiol.*, 9:88–93, 1999.

[CT91] T.M. Cover and J.A. Thomas. *Elements of information theory.* J. Wiley & Sons, New York, 1991.

[DA01] P. Dayan and L.F. Abbott. *Theorteical Neuroscience.* MIT Press, Cambridge, MA, 2001.

[DGM] P. Del Giudice and M. Mattia.

[Edg08] F.Y. Edgeworth. On the probable errors of frequency constants. *J. Roy. Stat. Soc. Ser. B*, 71:381–397, 499–512, 651–678, 1908.

[Edg09] F.Y. Edgeworth. On the probable errors of frequency constants. *J. Roy. Stat. Soc. Ser. B*, 72:81–90, 1909.

[ES97] C.W. Eurich and H. Schwegler. Coarse coding: calculation of the resolution achieved by a population of large receptive fields. *Biol. Cybern.*, 76:357–363, 1997.

[EW00] C.W. Eurich and S.D. Wilke. Multi-dimensional encoding strategy of spiking neurons. *Neural Comput.*, 12:1519–1529, 2000.

[Fer96] D. Ferster. Is neural noise just a nuisance? *Science*, 273(5283):1868–71, 1996.

[Fis22] R.A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London, Ser. A*, 222:309–368, 1922.

[Fit61] R. FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.*, 1:445–466, 1961.

[GG92] A. Gersho and R.M. Grey. *Vector quantization and signal compression.* Kluwer, Boston, 1992.

[GK02] W. Gerstner and W. Kistler. *Spiking Neuron Models. Single Neurons, Populations, Plasticity.* Cambridge University Press, Cambridge, MA, 2002.

[GM64] G.L. Gerstein and B. Mandelbrot. Random walk models for the spike activity of a single neuron. *Biophys. J.*, 4:41–68, 1964.

[GvH93] W. Gerstner and J.L. van Hemmen. Coherence and incoherence in a globally coupled ensemble of pulse-emitting units. *Phys. Rev. Lett.*, 71:312–315, 1993.

[H.84] Risken H. *The Fokker-Planck equation: methods of solution and applications.* Springer, Berlin, 1984.

[Haw71] A.G. Hawkes. Spectra of some self-exciting mutually exciting point processes. *Biometrika*, 58:83–90, 1971.

[Hel78] Hermann Helmholtz. The facts of perception. In R. Kahl, editor, *Selected Writings of Hermann Helmholtz.* Wesleyan University Press, Middletown, CT, 1878.

[Her] A.V.M. Herz. How is time represented in the brain?

[Her02] R. Herbrich. *Learning Kernel Classifiers.* MIT Press, Cambridge, MA, 2002.

[HH52] A.L. Hodgkin and A.F. Huxley. A quantitative description of ion currents and its applications to conduction and excitation in nerve membranes. *J. Physiol. (Lond.)*, 117:500–544, 1952.

[HHKM01] Sompolinsky H., Yoon H., Kang K., and Shamir M. Population coding in neuronal systems with correlated noise. *Phys. Rev. E*, 64:051904, 2001.

[Hin81] G.E. Hinton. Shape representation in parallel systems. In *Proceedings of the 7th Int. Joint Conference on Artificial Intelligence*, pages 1088–1096, Vancouver, BC, Canada, 1981.

[HMR86] G.E. Hinton, J.L. McClelland, and D.E. Rumelhart. Distributed representations. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing*, volume 1, pages 77–109. MIT Press, Cambridge, Massachusetts, 1986.

[HNT01] E. Haskell, D.Q. Nykamp, and D. Tranchina. Population density methods for large-scale modeling of neuronal networks with realistic synaptic kinetics: Cutting the dimension down to size. *Network: Comput. Neural Syst.*, 12:141–174, 2001.

[HS97] E.P Huang and C.F. Stevens. Estimating the distribution of synaptic reliabilities. *J. Neurophysiol.*, 78:2870–2880, 1997.

[Jay78] E.T. Jaynes. Where do we stand on maximum entropy inference. In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*. MIT Press, Cambridge, MA, 1978.

[Joh68] P.I.M. Johannesma. Diffusion models of the stochastic acticity of neurons. In E.R. Caianello, editor, *Neural Networks*, pages 116–144. Springer, Berlin, 1968.

[Joh96] D.H. Johnson. Point process models of single-neuron discharges. *J. Comput. Neurosci.*, 3:275–299, 1996.

[Jol86] I.T. Jollife. *Principal Component Analysis*. Springer, New York, 1986.

[KAM92] T.B. Kepler, L.F. Abbott, and E. Marder. Reduction of conductance-based neuron models. *Biol. Cybern.*, 66:381–387, 1992.

[Kar00] J. Karbowski. Fisher information and temporal correlations for spiking neurons with stochastic dynamics. *Phys. Rev. E*, 61:4235–4252, 2000.

[Kay93] S.M. Kay. *Fundamentals of Statistical Signal Processing, Vol. I (Estimation Theory)*. Prentice Hall, Upper Saddle River, NJ, 1993.

[KC02] L.C. Katz and J.C. Crowley. Development of cortical circuits: Lessons from ocular dominance columns. *Nature Rev. Neurosci.*, 3(1):34–42, 2002.

[KFK90] D.C. Knill, D. Field, and D. Kersten. Human discrimination of fractal images. *J.Opt.Soc.Am.A*, 7, 1990.

[Kni72] B. Knight. Dynamics of encoding in a population of neurons. *J.Gen.Physiol.*, 59, 1972.

[KWGL02] M. Kaschube, F. Wolf, T. Geisel, and S. Löwel. Genetic influence on quantitative features of neocortical architecture. *J. Neurosci.*, 22(16):7206–7217, 2002.

[KXFP01] C. Keysers, D. Xiao, P. Foldiak, and D. Perrett. The speed of sight. *J. Cog. Neurosci.*, 13:90–101, 2001.

[Lap07] L. Laplicque. Recherches quantiatives sur l'excitation electrique des nerfs traitee comme une polarization. *J. Physiol. Pathol. Gen.*, 9, 1907.

[LB96] W.B. Levy and R.A. Baxter. Energy efficient neural codes. *Neural Comput.*, 8:531–543, 1996.

[LBdRvS01] G.D. Lewen, W. Bialek, and R.R. de Ruyter van Steveninck. Neural coding of naturalistic motion stimuli. *Network: Comput. Neural Syst.*, 12:317–329, 2001.

[LC99] E.L. Lehmann and G. Casella. *Theory of point estimation.* Springer, New York, 1999.

[Lin86a] R. Linsker. From basic network principles to neural architecture: emergence of orientation columns. *PNAS*, 83(22):8779–83, 1986.

[Lin86b] R. Linsker. From basic network principles to neural architecture: emergence of orientation-selective cells. *PNAS*, 83(21):8390–8394, 1986.

[Lin86c] R. Linsker. From basic network principles to neural architecture: emergence of spatial-opponent cells. *PNAS*, 83(21):7508–7512, 1986.

[Lin88] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21:105–117, 1988.

[LO99] M.S. Lewicki and B.A. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A*, 16:1587–1601, 1999.

[LRS88] C. Lee, W.H. Rohrer, and D.L. Sparks. Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, 332:357–360, 1988.

[Mac02] C. Machens. *Sensory coding in natural environments: Lessons from the grasshopper auditory system.* PhD thesis, Humboldt-Universitaet zu Berlin, 2002.

[McI01] J.T. McIlwain. Population coding:a historical sketch. *Prog. Brain Res.*, 130:3–7, 2001.

[Mil96] K.D. Miller. Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns. In E. Domany, J.L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks*, volume 3, pages 55–78. Springer, New York, 1996.

[MKS89] K.D. Miller, J.B. Keller, and M.P. Stryker. Ocular dominance column development: Analysis and simulation. *Science*, 245:605–615, 1989.

[Mov99] J.A. Movshon. Deconstructing synchrony. *Invited Talk at the 13th Ann. Conference on Neural Information Processing Systems*, 1999.

[MS95] Z.F. Mainen and T.J. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268:1503–6, 1995.

[NB02] Fourcaud N. and N. Brunel. Dynamics of the firing probability of noisy integrate-and-fire neurons. *Neural Comput.*, 14:2057–2110, 2002.

[OF96] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:560–561, 1996.

[OF97] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Res.*, 37(23):3311–25, 1997.

[OKKS00] A. Omurtag, E. Kaplan, B. Knight, and L. Sirovich. A population dynamics approach to cortical columns with an application to orientation tuning. *Network: Comput. Neural Syst.*, 11:247–260, 2000.

[Par88] M.A. Paradiso. A theory for the use of visual orientation information which exploits the columnar structure of the striate cortex. *Biol. Cybern.*, 58:35–49, 1988.

[PBR$^+$96] S. Panzeri, G. Biella, E.T. Rolls, W.E. Skaggs, and A. Treves. Speed, noise, information and the graded nature of neuronal responses. *Network: Comp. in Neural Syst.*, 7:365–370, 1996.

[PDL01] A. Pouget, S. Deneve, and P.E. Latham. The relevance of fisher information for theories of cortical computation and attention. In J. Braun, C. Koch, and J.L. Davis, editors, *Visual attention and cortical circuits*, pages 265–283. MIT Press, Cambridge, MA, 2001.

[PFHD03] L. Paninski, M. Fellows, N. Hatsopoulos, and J. Donoghue. Spatiotemporal tuning properties for hand position and velocity in motor cortical neurons. *J. Neurophysiol.*, 2003.

[PTSR99] S. Panzeri, A. Treves, S. Schultz, and E.T Rolls. On decoding the responses of a population of neurons from short time windows. *Neural Comput.*, 11:1553–1577, 1999.

[PTT00] C.A. Parraga, T. Troscianko, and D.J. Tolhurst. The human visual system is optimised for processing the spatial information in natural visual images. *Curr Biol.*, 10(1):35–38, 2000.

[Rao46] C.R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1946.

[RKG62] R.W. Rodieck, N.Y.-S. Kiang, and G.L. Gerstein. Some quantitative methods for study of spontaneous activity of single neurons. *Biophys. J.*, 2:351–368, 1962.

[RMV01] D.S. Reich, F. Mechler, and J.D. Victor. Independent and redundant information in nearby cortical neurons. *Science*, 294:2566–2568, 2001.

[RS69] B.K. Roy and D.R. Smith. Analysis of the exponential decay model of the neuron showing frequency threshold effects. *Bull. Math. Biophys.*, 31, 1969.

[RT94] E.T. Rolls and M.J. Tovee. Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc. R. Soc. Lond. B*, 257:9–15, 1994.

[SA94] E. Salinas and L.F. Abbott. Vector reconstruction from firing rates. *J. Comput. Neurosci.*, 1:89–107, 1994.

[SBM$^+$04] G. Silberberg, M. Bethge, H. Markram, K. Pawelzik, and M. Tsodyks. Dynamics of population rate codes in ensembles of neocortical neurons. *J. Neurophysiol.*, 91, 2004.

[SBS$^+$87] B.C. Skottun, A. Bradley, G. Sclar, I. Ohzawa, and R.D. Freeman. The effects of contrast on visual orientation and spatial frequency discrimination: a comparison of single cells and behaviour. *J. Neurophysiol.*, 57:773–785, 1987.

[Sch89] H.G. Schuster. *Deterministic Chaos: An Introduction.* Wiley-VCH, New York, 1989.

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27:379–423 and 623–656, 1948.

[SK92] H.P. Snippe and J.J. Koenderink. Discrimination thresholds for channel-coded systems. *Biol. Cybern.*, 66:543–551, 1992.

[SK93] W.R. Softky and C. Koch. The hihgly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *J. Neurosci.*, 13:334–350, 1993.

[SKdRvSB98] S.P. Strong, R. Koberle, R.R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80:197–200, 1998.

[SLSZ03] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 2003. to appear.

[SN84] B. Sakmann and E. Neher. Patch clamp techniques for studying ionic channels in excitable membranes. *Annu. Rev. Physiol.*, 46:455–472, 1984.

[SN98] M.N. Shadlen and W.T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.*, 18:3870–3896, 1998.

[SO94] A. Stuart and K. Ord. *Kendall's advanced theory of statistics Vol. I.* Arnold, London, 1994.

[SO01] E.P. Simoncelli and B.A. Olshausen. Natural image statistics and neural representation. *Ann. Rev. Neurosci.*, 24:1193–1216, 2001.

[Sof96] W.R. Softky. Fine analog coding minimizes information transmission, 1996.

[SPPS04] E.P. Simoncelli, L. Paninski, J. Pillow, and O. Schwartz. Characterization of neural responses with stochastic stimuli. In M. Gazzaniga, editor, *The New Cognitive Neurosciences, 3rd edition.* MIT Press, Cambridge, MA, 2004. to appear.

[SS93] H.S. Seung and H. Sompolinsky. Simple models for reading neuronal population codes. *PNAS*, 90:10749–10753, 1993.

[SS02] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[SW49] C.E. Shannon and W.W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, IL, 1949.

[Swi80] NV Swindale. A model for the formation of ocular dominance stripes. *Proc. R.Soc. London B: Biological Sciences*, 208:243–264, 1980.

[Swi82] NV Swindale. A model for the formation of orientation columns. *Proc. R.Soc. London B: Biological Sciences*, 215:211–230, 1982.

[TFM96] S. Thorpe, D. Fize, and Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

[TKGA99] M. Tsodyks, T. Kenet, A. Grinvald, and A. Arieli. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286(5446):1943–6, 1999.

[TM91] F.E. Theunissen and J.P. Miller. Representation of sensory information in the cricket cercal sensory system. ii. information-theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *J. Neurophysiol.*, 66:1690–1703, 1991.

[TM95] F.E. Theunissen and J.P. Miller. Temporal encoding in nervous systems: a rigorous definition. *J. Comput. Neurosci.*, 2(2):149–162, 1995.

[TM97] T.W. Troyer and K.D. Miller. Physiological gain leads to high isi variability in a simple model of a cortical regular spiking cell. *Neural Comput.*, 9:971–983, 1997.

[Tod02] E. Todorov. Cosine tuning minimizes motor errors. *Neural Comput.*, 14:1233–1260, 2002.

[TS95] M. Tsodyks and T. Sejnowski. Rapid state switching in balanced cortical network models. *Network*, 6, 1995.

[Tuc88] H. Tuckwell. *Introduction to Theoretical Neurobiology.* Cambridge Univ. Press, Cambridge, MA, 1988.

[Vap95] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[vHvdS98] J.H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc R Soc Lond B Biol Sci.*, 265(1394):1724–1726, 1998.

[vNM47] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton Univ. Press, Princeton, 1947.

[Vog90] R. Vogels. Population coding of stimulus orientation by striate cortical cells. *Biol. Cybern.*, 64:25–31, 1990.

[vS96] C. vanVreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274, 1996.

[vV01] C. van Vreeswijk. Whence sparseness? In Volker Tresp Todd Leen, Tom Dietterich, editor, *Advances in Neural Information Processing Systems*, volume 13, Cambridge, MA, 2001. MIT Press.

[WC72] H. R. Wilson and J. D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, 12:1–24, 1972.

[WCF01] L.E. White, D. Coppola, and D. Fitzpatrick. The contribution of sensory experience to the maturation of orientation selectivity in ferret visual cortex. *Nature*, 411:1049–1052, 2001.

[WE01] S.D. Wilke and C. Eurich. Representational accuracy of stochastic neural populations. *Neural Comput.*, 14:155–189, 2001.

[WG98] F. Wolf and T. Geisel. Spontaneous pinwheel annihilation during visual development. *Nature*, 395:73–78, 1998.

[WNiA01] Si Wu, Hiroyuki Nakahara, and Shun ichi Amari. Population coding with correlation and an unfaithful model. *Neural Comput.*, 13(4):775–797, 2001.

[Zad98] A. Zador. Impact of synaptic unreliability on the information transmitted by spiking neurons. *J. Neurophysiol.*, 79:1219–1229, 1998.

[Zip49] G. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Cambridge, MA, 1949.

[ZS99] K. Zhang and T.J. Sejnowski. Neuronal tuning: to sharpen or broaden. *Neural Comput.*, 11:75–84, 1999.

Acknowledgment

It is true and not limited to brain research that the influence of the environment can hardly be overestimated and I feel that I will inevitably fail to acknowledge all those who have helped, in one way or another, to shape this thesis. First and foremost, however, I want to thank Klaus Pawelzik for his whole-hearted support, care, and friendship throughout the years. He not only gave me the freedom and means to explore and pursue my own research directions, but he was also there, whenever I needed help. I am also grateful to Professor Schwegler. I often got inspired by his thought-provoking comments and, although he was not my supervisor, I could feel his support from the back seat.

It is impossible to overlook the great efforts of David Rotermund who contributed a lot to the work on optimal coding by his amazing ability to let computers compute the 'uncomputable'. The steady support and helpfulness of Agnes Janssen, our office manager, and of Udo Ernst and David Rotermund, our voluntary system administrators, were exceptional. In addition, I want to mention my colleagues Bailu Si, Stefan Liehr, Roland Rothenstein, Stefan Wilke, Christian Eurich, Axel Etzold, Klaus Franke, Pit Hankel, Nadja Schinkel, Erich Schulzke, Andreas Thiel, Dennis Trenner, Ronald Bormann, Frank Emmert-Streib, and Rolf Henkel, who filled our group with life: there was always something to learn and discuss and there was always something to laugh about.

There are many researchers, not part of our group, who influenced my thoughts in a remarkable way. These are Matthias Kaschube, Fred Wolf, Karsten Kruse, Michael Herrmann, and Dirk Brockmann, who I met during my diploma work in the group of Theo Geisel at the Max-Planck-Institut für Strömungsforschung in Göttingen. I am also very grateful to Misha Tsodyks: several things, I got straight, because he was ready and patient to hear to my arguments, until it became clear, what my misunderstanding was. Furthermore, I want to mention Christian Machens, Jan Benda, Martin Nawrot, Christian Leibold, Richard Kempter, Bernhard Schölkopf, Matthias Franz, Michael Herzog, Andreas Kreiter, Lars Schwabe, and Jutta Kretzberg for light-hearted and inspiring meetings and discussions. I want to thank Rob de Ruyter van Steveninck, Andreas Herz, Leo van Hemmen, Wulfram Gerstner, and Martin Egelhaaf for their interest in my work and their support, and I want

to thank, in particular, Bruno Olshausen and Peter Dayan, who both played an important role during my search and decision process for a postdoc position. I am greatly indebted to Peter Schmiediche for his proof reading of my thesis, for which I actually let him almost no time.

Last but not least, I want to thank my whole family for their love and support, which is so all-embracing that I will not try to single out its particular meaning for this thesis. I will take the opportunity, however, to confess my deep thanks to my wife Christin, our two kids, Yara and Leon, as well as to my parents Holger and Regine, Ulrich and Annelie, and Walter and Roswitha.

Index

Publications

Articles

- G. Silberberg, M. Bethge, H. Markram, M. Tsodyks, and K. Pawelzik, Dynamics of population rate codes in ensembles of neocortical neurons, *under review.*
- M. Bethge, D. Rotermund, and K. Pawelzik. The influence of tuning width and dynamic range on the acuity of population codes, *to appear.*
- M. Bethge, D. Rotermund, and K. Pawelzik. A second order phase transition in neural rate coding: binary encoding is optimal for rapid signal transmission. *Phys.Rev.Lett.*, **90**: 088104, 2003.
- M. Bethge, D. Rotermund, and K. Pawelzik. Optimal neural rate coding leads to bimodal firing rate distributions. *Network: Comput. Neural Syst.*, **14**: 303-319, 2003.
- M. Bethge, D. Rotermund, and K. Pawelzik. Binary tuning is optimal for neural rate coding with high temporal resolution. *Advances in Neural Information Processing Systems*, **15** , to appear.
- M. Bethge, D. Rotermund, and K. Pawelzik. Optimal short-term population coding: when fisher information fails. *Neural Comput.*, **14(10)**: 2317-2351, 2002.
- M. Bethge and K. Pawelzik. Population coding with unreliable spikes. *Neurocomputing*, **44-46**: 323-328, 2002.
- J. Benda, M. Bethge, M. Henning, K. Pawelzik, and A.V.M. Herz. Spike-frequency adaptation: Phenomenological model and experimental tests. *Neurocomputing*, **38-40**: 105-110, 2001.
- M. Bethge and K.R. Pawelzik. Synchonous inhibition as a mechanism for unbiased selective gain control. *Neurocomputing*, **38-40**: 483-488, 2001.
- M. Bethge, K.R. Pawelzik, and T. Geisel. Brief pauses as signals for depressing synapses. *Neurocomputing*, **26-27**: 1-7, 1999.

Abstracts

- M. Bethge, D. Rotermund, and K. Pawelzik, Optimal neural population coding and Fisher information, in: *Verhandlungen der Deutschen Physikalischen Gesellschaft DPG (VI)* **37**, 1, eds. V. Häselbarth, Physik-Verlag GmbH, 509 (2002).

- M. Bethge, D. Rotermund, and K. Pawelzik, Optimal Short-Term Population Coding: When Fisher Information Fails, in: *Proceedings of the 28th Göttingen Neurobiology Conference*, eds. N. Elsner and G. W. Kreutzberg, Georg Thieme Verlag, Stuttgart , 250 (2001).

- M. Bethge, G. Silberberg, H. Markram, M. Tsodyks, and K. Pawelzik, Is population variance a signal for neocortical neurons?, in: *Proceedings of the 4th Meeting of the German Neuroscience Society 1*, eds. N. Elsner and G. W. Kreutzberg, Georg Thieme Verlag, 249 (2001).

- M. Bethge, G. Silberberg, H. Markram, M. Tsodyks, and K. Pawelzik, Realizing rapid population rate codes in ensembles of neocortical neurons, in: *Abstracts 265. WE-Heraeus-Seminar, Bad Honnef* (2001).

- M. Bethge and K. R. Pawelzik, A Physiological Model of Gain Control by Synchrony, in: *European Journal of Neuroscience* **12(11)**, eds. FENS Abstract, 490 (2000).

- M. Bethge and K. Pawelzik, Classification of coding paradigms with irregular spiking neurons, in: *Göttingen Neurobiology Report*, eds. Norbert Elsner and Ulf Eysel, Thieme , 896 (1999).

- M. Bethge, K. Pawelzik, and T. Geisel, Coherence detection with dynamic synapses, *Europ. J. Neursocience* **10(10)**: 22 (1998).

- M. Bethge, K. R. Pawelzik, and T. Geisel, Temporal coding and synfire chains in chaotic networks, in: *Proceedings of the 26th Göttingen Neurobiology Conference 2*, eds. N. Elsner and R. Wehner, Georg Thieme Verlag, 757 (1998).

- K. R. Pawelzik, M. Bethge, T. Geisel, and A. Kreiter, Irregular Spikes and Self-organized Synchronous Assemblies in a Cortical Microcircuit, in: *Society of Neuroscience Abstracts* **23(2)**, Society for Neuroscience, 1265 (1997).

Other puplications (German)

- M. Bethge and K. Pawelzik. Geheimsprache der Neuronen. *Gehirn & Geist*, **2**: 80-87, 2002. Spektrum.

- M. Bethge. Neuronale Kodierung mit dynamischen Synapsen. Diplomarbeit, 1998.

Lebenslauf

Name: Matthias Bethge

26.02.1973 Geboren in Wolfsburg als Sohn von Holger Bethge und Regine Bethge.

1979-1983 Besuch der Willhelm Busch Grundschule in Gifhorn.

1983-1985 Besuch der Erich Kästner Orientierungsstufe in Gifhorn.

1985-1992 Besuch des Humboldt Gymnasiums in Gifhorn.

1992 Abitur mit Auszeichnung am Humboldt Gymnasium in Gifhorn.

1992-1993 Zivildienst in Braunschweig.

1993-1998 Studium der Physik und Mathematik an der Georg-August Universität Göttingen.

1997-1998 Diplomand am Max-Planck Institut für Strömungsforschung in der Nonlinear Dynamics Group von Prof. Dr. T. Geisel.

1998 Diplom in Physik an der Georg-August Universität Göttingen.

1998-2003 Doktorand bei Prof. Dr. K. Pawelzik an der Universität Bremen, Mitarbeiter im Sonderforschungsbereich *Neurokognition* (SFB 517) der DFG.