TUM School of
Life Sciences

Efficient coding in populations of sensory neurons with multiple sources of noise and its relationship to phase transitions

Kai Röth

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Harald Luksch

Prüfende der Dissertation:
1. Prof. Julijana Gjorgjieva, Ph.D.
2. Prof. Dr. Ilona Grunwald Kadow

Die Dissertation wurde am 08.02.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 15.03.2021 angenommen.

Prior publication of parts of this thesis

Some of the methods, results, figures, tables, and text in this thesis are part of an article entitled *Efficient population coding depends on stimulus convergence and source of noise* which has been written together with Shuai Shao and Julijana Gjorgjieva. The article has been uploaded to the preprint server *bioRxiv* [1] and is currently under review for publication in the journal *PLOS Computational Biology*. All methods, results, figures, tables, and text from that article which are part of this thesis were my contribution to the article unless specifically mentioned otherwise.

Abstract

Sensory organs use neural populations to encode information about the environment and transmit it to the brain using parallel information channels. This happens in the presence of noise which corrupts the information encoding at several stages along the sensory pathways. The prominent hypothesis of efficient coding postulates that during evolution properties of sensory neurons have evolved to encode maximum information about stimuli given metabolic and biophysical constraints. Previous studies have used the efficient coding hypothesis to investigate small populations with just two neurons or only one noise source. They have shown that high noise levels favor redundant coding where the same signal is transmitted in parallel channels to average out noise, while low noise levels favor non-redundant coding where each channel transmits a distinct aspect of the signal.

In this thesis, I propose a model with populations with up to six neurons to derive the optimal encoding of a sensory signal in the presence of two different noise sources. I implement a model of spiking neurons jointly encoding a one-dimensional stimulus and maximize the mutual information between stimulus and population response using a constraint on the maximum firing rate. I also derive optimal thresholds of the neurons for varying sources and levels of noise assuming nonlinear input-output functions. In particular, the model includes input noise, which corrupts the signal before the nonlinearity and output noise which corrupts the signal after the nonlinearity. I determine critical noise levels where the optimal number of distinct neural thresholds changes, i.e. where the transition from redundant to non-redundant coding takes place. This is done for two scenarios: first, a lumped-coding channel where the information from different channels converges to a single channel, and second, an independent-coding channel when different channels contribute information without convergence. I find that the lumped-coding channel causes information loss, especially at intermediate noise levels, and therefore acts like an additional form of noise.

The transition from redundant to non-redundant coding with decreasing noise takes place gradually via subsequent bifurcations of optimal neural thresholds at critical noise levels. The thresholds bifurcate continuously with the independent-coding channel and discontinuously with the lumped-coding channel. Interestingly, there occurs an unexpected non-monotonic behavior of the optimal number of distinct thresholds for certain noise parameters.

I show that the threshold bifurcations at critical noise levels correspond to phase transitions encountered in physics and chemistry. In particular, maximizing the mutual information corresponds to minimizing the free energy of a physical system, noise corresponds to temperature, the threshold differences correspond to order parameters, and continuous threshold bifurcations correspond to continuous phase transitions.

To investigate how much information is lost in the case of suboptimal thresholds compared to optimal thresholds, I examine the shape of the information landscape as a function of the thresholds. I find that at continuous threshold bifurcations the information landscape takes the shape of a ridge, i.e. its curvature becomes zero in specific directions in threshold space, implying threshold combinations at which information does not change locally. Additionally, I quantify information loss in the case of randomly sampled thresholds or randomly perturbed optimal thresholds compared to optimal thresholds. My result suggests that it is not trivial to compare the information loss in the case of suboptimal thresholds across different neural population sizes. Nonetheless, using several measures that I develop, I demonstrate that the information loss with suboptimal thresholds compared to the case with optimal thresholds decreases with population size.

My results yield important insights into the coding strategies used by neural populations to optimally encode sensory stimuli in the presence of distinct sources of noise and can be applied to stimulus coding in diverse sensory systems.

Zusammenfassung

Sensorische Organe nutzen neuronale Populationen, um Informationen über die Umgebung zu kodieren und diese mittels paralleler Informationskanäle zum Gehirn weiterzuleiten. Dies passiert unter dem Einfluss von Rauschen, welches die Informationskodierung an verschiedenen Punkten des sensorischen Signalwegs stört. Die weit verbreitete Hypothese der effizienten Informationskodierung postuliert, dass sich im Laufe der Evolution die Eigenschaften der sensorischen Neuronen derart entwickelt haben, dass sie unter gegebenen metabolischen oder biophysikalischen Nebenbedingungen maximale Information über Stimuli kodieren. Bisher haben Studien die Hypothese der effizienten Informationskodierung genutzt, um kleine Populationen mit nur zwei Neuronen oder nur einer Rauschquelle zu untersuchen. Diese Studien haben gezeigt, dass hohe Rauschpegel eine redundante Kodierung begünstigen, bei der das gleiche Signal in parallelen Kanälen übertragen wird, um Rauschen herauszumitteln. Geringe Rauschpegel hingegen begünstigen eine nicht--redundante Kodierung, bei der jeder Kanal einen unterschiedlichen Aspekt des Signals überträgt.

In dieser Arbeit nutze ich ein Modell neuronaler Populationen mit bis zu sechs Neuronen, um die optimale Kodierung eines sensorischen Signals zu bestimmen, welches von zwei verschiedene Rauschquellen gestört wird. Dafür wähle ich ein Modell von feuernden Neuronen, in dem die Neuronen gemeinsam einen eindimensionalen Stimulus kodieren, und maximiere die Transinformation zwischen Stimulus und Populationsantwort unter der Nebenbedingung einer maximalen Feuerrate der Neuronen. Außerdem bestimme ich die optimalen Schwellenwerte der Neuronen bei verschiedenen Rauschquellen und -pegeln, wobei ich nichtlineare Input-Output-Funktionen annehme. Im speziellen enthält das Modell Inputrauschen, welches das Signal vor der Nichtlinearität stört, und Outputrauschen, welches das Signal nach der Nichtlinearität stört. Ich bestimme die kritischen Rauschpegel, bei denen sich die optimale Anzahl unterschiedlicher neuronaler Schwellenwerte ändert, das heißt, die Rauschpegel, bei denen der Übergang von redundanter zu nicht-redundanter Kodierung stattfindet. All dies führe ich unter zwei Szenarien durch: Zum einen mit einem konvergierten Kanal, bei dem die Information von verschiedenen Kanälen in einen einzelnen Kanal konvergiert. Zum anderen mit eigenständigen Kanälen ohne Konvergenz der Information aus den einzelnen Kanälen. Ich habe herausgefunden, dass die Konvergenz der Kanäle einen Informationsverlust verursacht, insbesondere bei mittleren Rauschpegeln und damit wie eine zusätzliche Rauschquelle wirkt.

Der Übergang vom redundanten zum nicht-redundanten Kodieren mit abnehmendem Rauschpegel findet graduell mittels aufeinanderfolgender Bifurkationen (Gabelungen) der optimalen neuronalen Schwellenwerte bei kritischen Rauschpegeln statt. Die Schwellenwerte gabeln sich auf stetige Weise mit den eigenständigen Kanälen und nicht-stetig mit den konvergierten Kanälen. Bei endlichen Rauschpegeln beeinflussen Input- und Outputrauschen die optimalen Schwellenwerte in einer ähnlichen Weise. Interessanterweise tritt bei bestimmten Rauschparametern ein unerwartetes nicht-monotones Verhalten der optimalen Anzahl unterschiedlicher Schwellenwerte auf.

Die Schwellwertbifurkationen bei kritischen Rauschpegeln entsprechen physikalischen Phasenübergängen. Insbesondere entspricht die Maximierung der Transinformation der Minimierung der freien Energie eines physikalischen Systems, das Rauschen entspricht der Temperatur, die Schwellwertdifferenzen entsprechen Ordnungsparametern und stetige Schwellwertbifurkationen entsprechen stetigen Phasenübergängen.

Um zu untersuchen, wie viel Information im Falle von suboptimalen statt optimalen Schwellenwerten verloren geht, untersuche ich die Form der Informationslandschaft in Abhängigkeit der Schwellenwerte. Ich habe herausgefunden, dass bei stetigen Schwellwertbifurkationen die Informationslandschaft die Form eines Grats annimmt, das heißt, die Krümmung in bestimmten Richtungen des Schwellwertraums ist null, was Schwellwertkombinationen impliziert, bei denen die Information sich lokal nicht ändert. Außerdem quantifiziere ich den Informationsverlust im Falle von zufällig gewählten Schwellenwerten oder Störungen von zufälliger Stärke der optimalen Schwellenwerte vergleichen mit optimalen Schwellenwerten. Das Ergebnis ist, dass es nicht trivial ist, den Informationsverlust durch suboptimale Schwellenwerte über die Populationsgrößen hinweg zu vergleichen. Ich entwickele verschiedene Maße, um trotzdem zeigen zu können, dass der relative Informationsverlust durch suboptimale Schwellenwerte mit der Populationsgröße abnimmt.

Meine Ergebnisse liefern wichtige Erkenntnisse bezüglich der optimalen Strategien von neuronalen Populationen, um sensorische Stimuli unter dem Einfluss von verschiedenen Rauschquellen zu kodieren. Diese Ergebnisse können auf die sensorische Kodierung in verschiedenen sensorischen Systemen übertragen werden.

Contents

List of Figures

List of Tables

Acronyms

ANF	auditory nerve fiber
BFGS	Broyden–Fletcher–Goldfarb–Shanno (sic)
CMP	Conway-Maxwell-Poisson
GND	generalized normal distribution
LNP	linear-nonlinear model with a probabilistic spike generation process
NM	Nelder-Mead
MI	mutual information
MSE	mean square error
RGC	retinal ganglion cell
SLSP	sequential least squares programming

1 Subject and motivation

Where do we come from? What are we? Where are we going? Both philosophy and natural sciences are originally driven by such deep questions. Today we know that living and reproducing organisms have appeared at some point in planetary history and then biological evolution with its “survival of the fittest” has led to an explosion of variability in living species. For all organisms, it is essential to interact with their respective environments and in many animals, specific cells and organs have evolved to receive and process information about their surroundings [2]. These cells and organs are often highly specialized for different sensory modalities like light [3], sound [4], chemical substances [5], forces [6], temperature [7], mechanical vibrations [8], and electrical [9] or magnetic fields [10]. Through evolution, many of them have reached an astonishing degree of sensitivity, precision, and complexity.

However, interacting with the environment is not just about sensors collecting information about the surroundings but also about processing the information and making use of it through prompting behavior which is – ideally – in the interest of the respective animal or its species. To achieve this, extremely complex organs like brains have evolved which integrate information from different sensory modalities. The brain is so complex that science is still very far away from understanding its function and many scientists argue that it will never be possible to understand it at all.[1] Thus, it is a commonly held opinion that to understand “the brain” one should first understand simpler neural circuits where the complexity is smaller and where the input and output and the immediate purpose of the circuits are better known [12,13]. At least the part about having more knowledge of input/output and purposes is true for sensory organs. Therefore, it makes sense to investigate the basic principles of how sensory organs work and use these insights to understand how our brains work.

A very important building block of sensory organs are neurons, i.e. electrically excitable cells specifically evolved for information processing. They are well-suited for information processing since they are fast and already a single neuron can perform basic computations like addition and multiplication [14]. When they are coupled together via *synapses* the resulting circuits can in many cases perform highly complex computations [15]. The example of a “simple” animal, the fly, demonstrates that modern computers are still far away from reaching similar performance in many sensory tasks as, for instance, object movement detection – in particular when it

[1] This is, for example, expressed in the famous statement “If the human brain were so simple that we could understand it, we would be so simple that we couldn’t.” [11]

comes to energy efficiency [16].

With my work, I want to contribute to the understanding of sensory processing and why sensory organs have evolved the specific way they have. In particular, the goal is to better understand basic principles and phenomena which are present across different sensory organs and different species. Across different species, the same sensory organ, for example, the eye, often seems to have very differently evolved [17]. However, in most species, it still shares many commonalities when it comes to its basic structure, building blocks, basic processes, and purpose.

At the same time, every sensory organ is composed of a large variety of different neuron types, which differ in size, shape, response properties, function, and gene expression patterns, to just name a few [18, 19]. My goal is to shine a light on why exactly it made sense from an evolutionary perspective that cells in sensory organs show such a great diversity. To achieve this, I use an approach from an information theoretical perspective: information from sensory organs has to be sent to the brain. However, the amount of information that can be transmitted per unit time is limited, for what often the metaphor of the "information bottleneck" [20] is used (Fig. 1.1A) . Moreover, information transmission using neurons needs substantial amounts of energy [21]. A prominent hypothesis called *efficient coding* argues that during evolution the energy efficiency of information transmission has been optimized [22, 23] (Fig. 1.1B). The reason is that animals that could transmit more information per unit energy had an advantage in spotting predators or prey, could better communicate with conspecifics, or needed less food compared to fellow animals, all of that meaning higher evolutionary fitness.

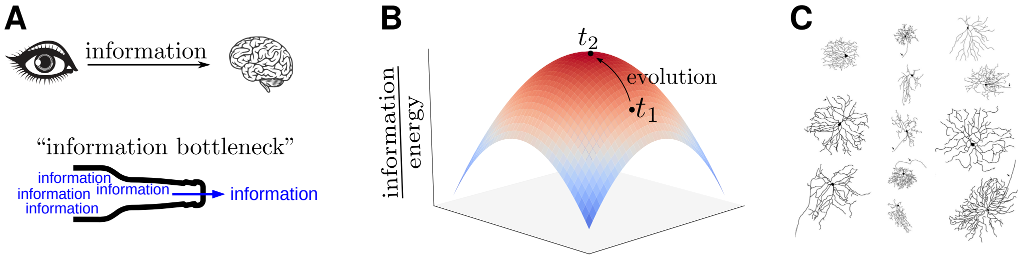


Figure 1.1: Diversity of sensory cells and efficient information transmission. **A.** Information about the environment is obtained by sensory organs and sent to the brain. However, the amount of information that can be transmitted is limited. **B.** Schematic depicting the assumption that due to evolutionary pressure information transmission has become close to optimal. If at some point in time (t_1) the efficiency of information transmission for a biological species is suboptimal, then over time it evolves towards being optimal optimal (t_2). Optimality here means maximally efficient in terms of information transmission per energy needed. **C.** Retinal ganglion cells, the output cells of the eye, show a large variability (here: morphology of dendritic trees). From [24].

Retinal ganglion cells (RGCs) are the neurons that send information from the eye to the brain [18]. Figure 1.1C shows how RGCs from the rat retina differ in size and shape. The goal of my work is to use the efficient coding hypothesis to gain an understanding of why it might be favorable to have such a diversity of neurons in sensory organs. My approach is not restricted to a specific sensory organ but can be applied to any sensory modality.

In all systems in which information is processed or transmitted, noise limits the amount of information that can be processed [25]. This is especially true for biological systems, where noise corrupts the information processing at many different stages [26]. Nevertheless, many sensory organs function astonishingly well in surroundings dominated by noise as for example star light [27] or cocktails parties [28]. The efficient coding hypothesis can also be used to predict optimal information processing in the presence of noise. In this thesis, I will use it to investigate how optimized information encoding depends on the type and strength of noise. Interestingly, varying noise can cause sudden changes in the way sensory signals are optimally encoded and it has been shown that these sudden changes represent phase transitions from physics [29]. In physics, phase transitions occur when the properties of a system suddenly change with an external variable like temperature, for example in the transition from water to ice. Therefore, I am also going to characterize the properties of the phase transitions occurring in my studies.

The understandings gained about information encoding in sensory organs might be beneficial in understanding diseases of sensory organs and could aid the development of medical technology that helps patients. Widely-known examples are retinal and cochlear implants [30,31]. However, also other approaches like gene therapy or optogenetic therapy for sensory diseases are under active development [32]. Apart from a more general understanding of how information is encoded, it might also help to understand the role of different cell types and how therapies could be specifically tailored towards certain cell types.

This thesis is organized as follows: In Chapter 2, I give a basic introduction into how sensory neurons are modeled, the basics of information theory, and how sensory organs use different neuron types to transmit information in parallel. This is followed by a literature overview of studies using the efficient coding hypothesis and how noise influences optimal information encoding. In Chapter 3, I describe the model that I use for my studies and present results of optimal neural diversity when maximizing information transmission. It will become apparent, that optimal diversity changes at particular noise levels. In Chapter 4, I discuss the shape of the information landscape and why it takes a specific form at particular noise levels. This includes quantifying information losses in the case of suboptimal information processing. Furthermore, I show that the sudden changes in optimal neural diversity can be related to phase transitions. Finally, I will discuss my results and the limits of my model in a greater context (Ch. 5).

2 Foundations and literature

To investigate how efficient information encoding has led to the development of different neuron types in sensory organs, in this chapter, I first give an introduction into the necessary concepts. This includes a basic neuron model used for sensory neurons (Sec. 2.1), how sensory systems encode stimuli with parallel information streams (Sec. 2.2), and an introduction into information theory (Sec. 2.3). Afterward, I give an overview of the literature about how previous studies have made use of the efficient coding hypothesis (Sec. 2.4), including the properties of noise and its effects on efficient information encoding, and the convergence of parallel channels. Finally, I will give a basic introduction to phase transitions (Sec. 2.5).

2.1 The linear-nonlinear spiking neuron model

The knowledge presented in this section is from [14] unless stated otherwise. Neurons are electrically excitable cells that play a crucial role in information processing. They have a *membrane potential*, i.e. a voltage difference between the inner and the outer side of their cell membrane. Inputs to a neuron can either decrease or increase the membrane potential. Most neurons are spiking neurons, which means that if the membrane potential increases above a certain threshold value the neuron responds with a characteristic output pulse, the *spike.* That is why input that increases the membrane potential is called *excitatory* input while input that decreases the membrane potential is called *inhibitory* input. If the threshold value of the membrane potential is not reached no output is generated. Thus, neurons are highly nonlinear units.

The *Hodgkin-Huxley model* is a mathematical model that describes the dynamics of channel openings and closings depending on the membrane potential [33]. It can explain in great detail how the membrane potential changes with incoming inputs and how the temporal dynamics evolve if the threshold is crossed. However, I take a less detailed level of description in my work. To model the input-output function of sensory neurons, it is common to use a linear-nonlinear model with a probabilistic spike generation process, or in short *LNP* model. There, the input is in general a time-dependent sensory stimulus that is first processed by a linear function, then passed through a nonlinear function, and the output of the nonlinear function is finally used to generate stochastically distributed spikes (Fig. 2.1). The sequence of the generated spikes is called the *spike train* and it encodes information about the sensory stimulus. Such a mapping from stimulus to spike train can be implemented by a variety of sensory systems, for instance, the retina which processes various visual stimulus attributes, such as light intensity or contrast [34], the olfactory receptor neurons which process a range of concentrations of odor molecules [5, 35–37], or the auditory nerve fibers (ANFs) which transmit information about sound pressure levels [4]. I will now briefly describe these three steps.

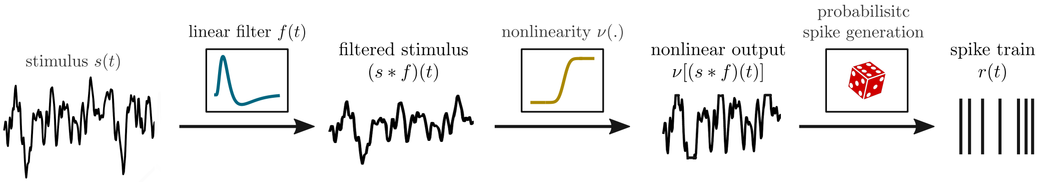


Figure 2.1: Modeling sensory neurons with a linear-nonlinear model with a probabilistic spike generation process. From left to right: the input to the neuron, a time-dependent stimulus $s(t)$, is first convolved by a linear filter $f(t)$, whose output $s * f(t)$ is then passed through the nonlinear function $\nu(.)$. Its output is then used as a rate variable for the stochastic spike generation process. The resulting spike train $r(t)$ carries information about the stimulus $s(t)$. Based on [14].

2.1.1 Linear filters as a model of the receptive fields of sensory neurons

A linear filter is used to model the *receptive field* of a sensory neuron. The receptive field describes to which spatial and temporal stimulus characteristics (also called *features*) the neuron responds. In the visual system, spatial features represent the location of stimulus in the visual field and spatial changes of the stimulus as for example edges or bars of certain orientations. Temporal features are for example the onset and offset of a stimulus. Mathematically, the process of linear filtering is described as a convolution of the filter *kernel*, $f(t)$, with the stimulus, $s(t)$, so that the output of the linear filter is[1]

$$L(t) = (s * f)(t) \equiv \int_0^\infty f(\tau)s(t-\tau)\mathrm{d}\tau \ . \tag{2.1}$$

In biology, the filter $f(t)$ is approximated by stimulating the cell with uncorrelated inputs ("white noise") for some time and determining the *spike-triggered average* of the stimulus, i.e. the average shape of the stimulus in the time before a spike is emitted by the neuron. $f(t)$ often has a biphasic shape, meaning a positive part followed by a negative part. This shape is responsible for the adapting behavior of the neural output with constant stimulus: a temporal biphasic filter promotes stimuli changing in time but is unresponsive to constant stimuli, while a spatial biphasic filter promotes edges and spatial structure but is unresponsive to full-field stimulation.

[1] Here, I just express the temporal dimension of the linear filter, but the spatial dimensions (of the retina) or the frequency dimension (in the cochlea) are completely analogous.

There exist several studies who have investigated the optimal shapes of linear filters [38–42]. The work of this thesis will be dedicated to the nonlinear function and the spike generation. Even though the linear filter should in principle have an influence on the nonlinear processing, its effect can be incorporated easily by a change in the input distribution. Thus the framework used in the following chapters will be a general framework independent of the particular linear filtering. For simplicity I set $s(t) \equiv L(t)$ and denote $s(t)$ as the linearly filtered or linearly *preprocessed* stimulus from now on.

2.1.2 Tuning curves as the nonlinear functions

The nonlinear part of the LNP model is called the *tuning curve* and describes the nonlinear response characteristics of a neuron. It is given as a function of the filtered stimulus $s(t)$ and the firing rate $\nu(t)$. Experimentally, the nonlinearity is obtained by plotting $\nu(s)$ (obtained from the spike train, see next section) at different s and fitting a suitable function. In principle, different shapes of the function are possible. Sensory neurons are often modeled by using a sigmoidal ("S-shaped") function [14, 29, 43, 44] (Fig 2.2A). Qualitatively, they show the following characteristics: for small input values s, the firing rate is equal to the baseline firing rate, also called "spontaneous" firing rate, ν_0. The input value for which the firing rate starts to be significantly larger than ν_0 is called the threshold, θ. From there on, the firing rate increases until the maximum firing rate, $\nu_{\max}$ is reached. There, even with a further increase of the input, the firing rate does not increase anymore. Such a function is highly nonlinear since the firing rate does not change at small or large input values (where the function is flat), but changes a lot in an intermediate regime (where the function is steep). The steepness of the function is related to the sensitivity of the neuron since small input changes cause large output changes. Figure 2.2B shows tuning curves of different ANFs, which encode sound pressure levels. All of them have a sigmoidal shape but they differ in spontaneous rate, threshold and maximum firing rate. The goal of this thesis is to contribute in understanding the reasons for this diversity in tuning curves.

2.1.3 The probabilistic spike generation process

For a given input, neurons in general do not produce deterministic output. In the LNP model, the linear and the nonlinear parts are deterministic and the stochastic nature of neural output is incorporated by a probabilistic spike generation process. That means that the output of the nonlinearity, $\nu(t)$, indicates how likely it is that spikes are emitted, however, the exact spike times are generated by a stochastic process and are thus randomly distributed. The most common stochastic process used in the LNP model is the *Poisson process*.[2] It takes the rate ν as input and is characterized via the Poisson distribution in such a way that the number of spikes k

[2] The Poisson process is in fact so common that the acronym LNP is often understood to be meant linear-nonlinear *Poisson* [14, 46].

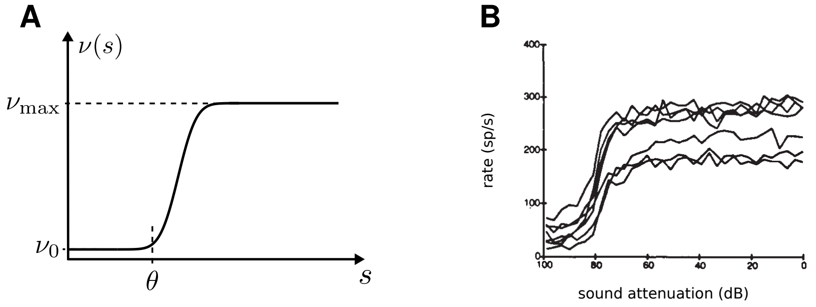


Figure 2.2: The nonlinearity of sensory neurons with a monotonic tuning curve can be modeled using a sigmoidal function. A. Basic properties of a tuning curve modeled with a sigmoidal function: the firing rate ν depends on the (linearly filtered) input stimulus s. Below a threshold θ, the firing rate is the baseline (or spontaneous) firing rate ν_0. The maximum firing rate is denoted by $\nu_{\max}$. **B.** Measured tuning curves from seven different auditory nerve fibers. From [45].

in a time interval ΔT follows the Poisson distribution:

$$P(k|\nu) = \frac{(\nu \Delta T)^k}{k!\ e^{-\nu \Delta T}}\ . \tag{2.2}$$

Strictly speaking, this is true only for a constant rate ν or the time interval approaching zero and this special case is called the *homogeneous* Poisson process. Since the firing rate of a neuron changes in time, the *inhomogeneous* Poisson process is the way to go. There, the number of spikes $k(t_a, t_b)$ in a time interval $[t_a, t_b]$ is given as [47]

$$P(k(t_a, t_b)|\Lambda) = \frac{\Lambda^k}{k!\ e^{-\Lambda}}\ , \tag{2.3}$$

where Λ is the integrated rate in$[t_a, t_b]$:

$$\Lambda := \int_{t_a}^{t_b} \nu(t) \mathrm{d}t\ . \tag{2.4}$$

The special property of a Poisson process is that spikes are generated independently of each other, i.e. the fact of a spike generation at time t_1 has no influence on the probability of another spike generation before or after t_1. Apart from being biologically unrealistic for very short[3] time intervals around t_1, it describes the stochastic behavior of many sensory neurons reasonably well. In my work, I almost exclusively use the Poisson process for modeling the spike generation process, except for

[3] The *refractory period* does not allow another spike generation for 1 to 2 ms after a previous spike and reduces the probability of spike generation for another few milliseconds. In principle, this could be taken into account by the model by setting the firing rate to zero for 2 ms after a spike and decreasing it by some factor for another few milliseconds.

Section 3.6.4.2, where I study two other processes.

Note: The assumption is that sensory neurons which transmit information about stimuli to the brain (e.g. RGCs in the visual system and ANFs in the auditory system) encode the information with their spiking outputs. Therefore, I use the words information *transmission* and information *encoding* as synonyms throughout my work when I talk about sensory neurons.

2.2 Sensory systems encode stimuli with parallel information streams

In many sensory systems, the sensory signal is not just coded by individual neurons but rather by the joint activity of populations of neurons. One signature of this parallel coding might be the remarkably diverse response properties exhibited by many sensory neurons. Figure 2.3 illustrates this parallel encoding using the eye as an example: the visual image is mapped onto the retina and each point of the retina is covered by the receptive fields of different RGC types which each transmit a different version of the image to the brain [48, 49]. In specific, it has been shown that around thirty different RGC types exist [18]. For each type, the receptive fields independently tile the retina with little overlap, meaning each spot of the retina is covered once and only once by each RGC type (see Fig. 2.3A for a schematic showing the receptive fields of three different RGC types). The RGC types differ in size, morphology, connectivity, and response properties [18, 50]. Several of them show very specific response properties like direction selectivity or object movement selectivity, while for other RGC types, the exact properties still remain elusive. Altogether, it can be said that many RGC types encode different visual features of the same retinal image [18, 49, 51] (Fig. 2.3B,C). Yet, there are also neurons from distinct RGC types which in parallel encode a single stimulus feature but differ in functional properties like thresholds [29, 52, 53]. Altogether, this means that the information stream from the eye to the brain is highly parallelized (Fig. 2.3D).

Another example of parallelization in a sensory information stream is the first synapse level of the auditory pathway, where each inner hair cell transmits information about sound intensity to approximately ten to thirty different ANFs [55]. ANFs differ in several aspects of their responses, including spontaneous rates and thresholds [56]. However, each fiber receives exclusive input from only a single inner hair cell. As in the retina, this results in a highly parallelized stream of sensory information. Similarly, this parallel encoding of a single stimulus feature with a population of neurons with different thresholds has been shown in olfactory receptor neurons [5], in mammalian touch receptors [57], and in electroreceptors of electric fish [9].

In conclusion, one can say that the remarkably distinct functional properties of

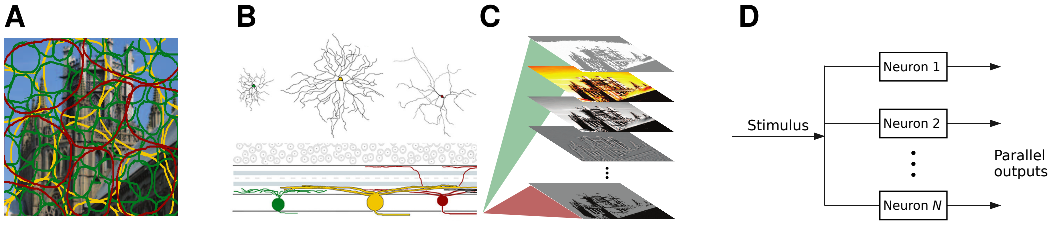


Figure 2.3: Sensory systems encode stimuli with parallel information streams. Schematic illustrating how different neuron types in the eye transmit different features of the visual input in parallel. **A.** The whole retinal image is covered by distinct retinal ganglion cells (RGCs). Each of the three RGC types depicted here (green, yellow, red; they are unrelated to the three primary colors) independently tiles the retina with the receptive fields (circles) of the respective cells. Thus in this schematic, each point of the retinal image is covered exactly once by each of the three distinct RGC types. Redrawn from [49]. **B.** Top: the dendritic trees of three RGC types. Note that they are artificially separated in space for visualization purposes but in reality, dendritic trees of different RGC types overlap. Bottom: schematic slice through the retina. The axons of the RGCs (lines at the very bottom) transmit the information towards the brain. Schematic from [54]. **C.** Each RGC type transmits a different representation or a different feature of the same retinal image to the brain. This results in parallel information streams. **D.** Schematic illustrating that the same stimulus is encoded in parallel by several neurons.

sensory neurons are directly related to the parallel nature of the information stream from sensory organs to the brain. Since the goal of this thesis is to explain diversity among sensory neurons using an information theoretical approach, the next section addresses the foundations of information theory.

2.3 Information theory

The central topic of my thesis is information encoding by sensory organs. For that, I need a definition of information and a framework that can quantify it. Claude Shannon's groundbreaking work in the 1940s introduced information theory and he defined measures which are very useful for quantifying how much information is encoded [25]. In this section, I will describe the basic concepts of *information entropy* and *mutual information* and how they can be applied to studying information encoding with neural populations. If not stated otherwise, my sources of knowledge for this section are [14, 25].

The basic concept of information theory is a communication *channel* that is supposed to transmit information from the *sender* (in my work: sensory neurons) to the *receiver* (the brain). The sender uses certain *symbols* (neural responses) to encode information about the *source* (the stimulus value) that is supposed to be transmitted. Let us denote the stimulus with S and the neural responses with R. Both are random variables, i.e. they follow probability distributions $p(S)$ and $p(R)$, respectively. To quantify the amount of information that a sender can transmit to the receiver, Shannon introduced the concept of *entropy*. It is very related to entropy known from physics but in information theory, it quantifies how "surprising" a random variable is on average. If the response value r occurs with probability $p(r)$ then the surprise of the occurrence of r should be quantified by $-\log p(r)$ according to Shannon. This makes sense for the following reason: if $p(r)$ is very small, it is very surprising if r occurs; while when $p(r)$ is very large it occurs all the time and nobody is surprised if it occurs.[4] The entropy $H(R)$ of a random variable R quantifies how surprising R is on average:

$$H(R) = -\sum_{r \in R} p(r) \log_2 p(r) \ . \tag{2.5}$$

Its unit is denoted as *bits* and $H(R) \geq 0$. The entropy takes larger (maximum) values when all r appear with similar (equal) probabilities. $H(R)$ also becomes larger when R can take more values: the surprisingness for each r is increased when they appear less often and the sum is taken over more terms. If R, for example, can take two values then $H(R)$ can be at most 1 bit – namely in the case when $p(r_1) = p(r_2) = 1/2$. If R can take four different values, then $H(R)$ can take at most 2 bits (when $p(r) = 1/4$ for all r).

Information theory states that the information is bounded by the entropy of the variable that the sender uses to transmit information to the receiver (here: $H(R)$). However, at the same time, information channels are in general corrupted by noise, which further reduces the amount of information. The concept of noise can be understood as an imperfect correlation between stimulus and response: if stimulus and response are perfectly correlated, then the response is fully determined by the stimulus and there is perfect information transmission. If, on the other hand, the response is independent of the stimulus, then the response cannot carry *any* information about the stimulus. Shannon introduced the *noise entropy* to quantify the amount of noise. The noise entropy is given as the entropy of the response for a given stimulus value, averaged over all stimulus values s:

$$\begin{aligned} H(R|S) &= \sum_{s \in S} p(s) H(R|s) && (2.6)\\ &= -\sum_{s \in S} p(s) \sum_{r \in R} p(r|s) \log_2 p(r|s) \ . && (2.7) \end{aligned}$$

This makes intuitive sense since for a given stimulus the surprise of the response should ideally be zero, i.e. if for a given stimulus there is a surprise about a response

[4] There are two reasons why Shannon chose the negative logarithm. First, the surprise of r should increase with decreasing $p(r)$, and second, surprise should add up, i.e. the surprise of observing both r_1 and r_2 should be the surprise of observing r_1 plus the surprise of observing r_2. Since $p(r_1, r_2) = p(r_1)p(r_2)$ (if r_1 and r_2 are independent), the logarithm is chosen because it allows to transform the product into a sum.

then something is not ideal in the information encoding: noise has struck. Shannon showed that the amount of information that the sender transmits to the receiver is given as the difference between response entropy and noise entropy. He termed this quantity *mutual information*, I_m:

$$\begin{aligned} I_m(R;S) &= H(R) - H(R|S) && (2.8) \\ &= -\sum_{r\in R} p(r)\log_2 p(r) + \sum_{r\in R}\sum_{s\in S} p(s)p(r|s)\log_2 p(r|s) && (2.9) \\ &= \sum_{r\in R}\sum_{s\in S} p(s)p(r|s)\log_2 \frac{p(r|s)}{p(r)} \ . && (2.10) \end{aligned}$$

In the case of zero noise, $I_m(R;S) = H(R)$; one says "the channel reaches capacity".[5] Furthermore, $I_m(R;S) = 0$ in the case of infinite noise – this corresponds to zero correlation between S and R. Finally, the mutual information is symmetric, i.e. $I_m(R;S) = I_m(S;R)$. In conclusion, the mutual information is a measure of how much one can know on average about one variable when observing the other.

Note that by summing over each r and s, I have implied that both S and R are discrete variables. This makes sense when using a spike count for the neural response. The stimulus, however, is in general continuous. Thus, I will replace the sum over s by an integral over s from now on. Furthermore, for a population of neurons, one has a spike count for each neuron, such that the overall neural response is a vector. This is also easily captured by the mutual information:

$$I_m(\vec{R};S) = \sum_{\vec{r}\in\vec{R}} \int_s p(s)p(\vec{r}|s)\log_2 \frac{p(\vec{r}|s)}{p(r)} \mathrm{d}s \ . \qquad (2.11)$$

Using a population of neurons, the number of response states increases exponentially with the number of neurons. Therefore, using parallel information streams with populations of neurons can in principle drastically increase the amount of information that can be encoded.

Also other measures exist which can quantify information encoding, for example, the mean square error between the stimulus and the estimated stimulus. These measures, however, make specific assumptions about a decoding mechanism, i.e. a mechanism where the stimulus is estimated from the neural response. The mutual information is a more general approach since it gives an upper bound for the information that can be extracted by any decoder [58]. Another quantity is the *Fisher*

[5] Note that in this case of zero noise, the information is still bounded by the alphabet of the response variable: with an alphabet of M distinct responses and no noise, the maximum information is $\log_2 M$ bits – namely when all response states occur with probability $p = 1/M$. For unbounded information, the response variable would need to have infinitely many distinct states. This is only realistic with a continuous response variable, which, however, can never be perfectly correlated with the stimulus variable in real scenarios. Thus the capacity of every information channel is limited.

information, which quantifies the sensitivity of tuning curves among the stimulus axis, but is not suited for spiking neurons with considerable noise [59, 60].

2.4 The efficient coding hypothesis

The question I want to answer in this thesis is whether the diverse properties of sensory neurons are a consequence of the evolutionary pressure of the sensory system to efficiently encode sensory stimuli. A powerful framework to address this question is the hypothesis of efficient coding [22]. Both obtaining more information about the environment and having less metabolic costs[6] are beneficial from an evolutionary perspective. The efficient coding hypothesis postulates that during evolution the efficiency of information encoding in sensory systems has been optimized. Information efficiency can be increased by increasing the amount of information or by decreasing the metabolic cost of information encoding. This means there is a so-called *Pareto optimality* where optimal information efficiency can be reached at any given metabolic budget with resulting different amounts of information [61]. To give an example, an elephant and a fly have a very different metabolic budget, however, both can maximize information transmission for their respective metabolic budget. This illustrates that information efficiency is optimized with respect to constraints. Such a constraint can be the metabolic budget, but also the available space and material as well as numerous biophysical and phylogenetic constraints [2]. Table 2.1 lists some of the most important limits and constraints.

Table 2.1: List of constraints in neural coding. Data from [2].

Type of limit	**Specific limit**	**What is constrained; Examples**
finite Resources	space	size of the eye / optic nerve / skull
	material	dendritic tree: minimizing cable length
	energy	firing rates, axon diameters, synapse maintenance
biophysical limits	noise	min. number of cells / spikes / ion channels / ...
	membrane capacitance	max. firing rate /cell size
	cytoplasm conductivity	max. firing rate / cell size
	energy density	max. and mean firing rate
phylogenetic limits	genetic building blocks	limited toolbox: ion channels, pumps, receptors, ...
	developmental history	retina is “inside out”: occurrence of blind spot

Predictions from efficient coding are consistent with many properties of primary sensory neurons. The presence of center-surround receptive fields in the visual pathway reduces unnecessary redundancy and thus increases the efficiency of information transmission [38], and this has also been shown for the split into ON and OFF pathways[7] in the retina [41, 62]. Similarly, the division of RGCs into

[6] “Metabolic cost” is a fancy term for “requires some amount of energy through food”.

[7] ON cells are neurons which increase their response with increasing stimulus, while OFF cells are neurons which decrease their response with decreasing stimulus.

midget and parasol cells[8] in the primate retina can be explained by an efficiency gain [63, 64]. The efficient coding theory also predicts an alignment of the stimulus statistics and the input-output function of sensory neurons. This has been shown to be the case for the retinal cells of the blowfly [43] and ANFs from the bullfrog [65]. Furthermore, it has been shown in movement-selective cells in the insect retina [44] and in neural populations in the auditory brain stem [66] that the input-output functions adapt to changing stimulus statistics in order to maintain maximized information transmission in different environments.

2.4.1 Noise corrupts information transmission in sensory systems

In information theory, noise is defined as an unwanted disturbance of a signal and reduces the information capacity of a communication channel [25]. In biological information processing, noise is an ubiquitous phenomenon that corrupts signal transmission at different processing stages [26] (Fig. 2.4).

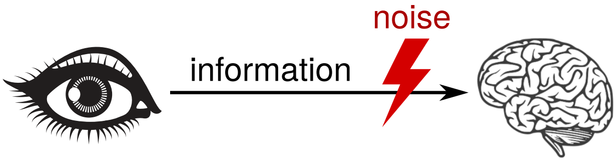


Figure 2.4: Noise negatively impacts information transmission.

Neural information transmission is based on moving chemicals or ions. All movements of molecules or ions are affected by *Brownian motion* – i.e. random thermal fluctuations – especially when molecule or ion concentrations are low or transmission distances are large [2]. In addition to this Brownian noise, the opening and closing of ion channels have randomness involved. This *channel noise* [67] leads to membrane potential fluctuations which are especially impactful when the membrane potential is near the threshold for spiking. Furthermore, noise is introduced at synapses. This *synaptic noise* is caused by spontaneous opening of intracellular calcium stores, calcium channel noise, or spontaneous fusion of neurotransmitter vesicles with the cell membrane [26]. Apart from this general noise in neurons, there exist also noise sources in the early sensory pathways. Every sensory modality is subject to *transducer noise* [68], meaning that no conversion from physical quantities to a neural can be perfectly reliable. In the visual modality, for example, *photon noise* exists – photons arrive Poisson distributed, which implies noise in low illumination settings [69, 70]. Furthermore, the light-sensitive molecules – the *opsins* – are subject to spontaneous thermal isomerization, causing false positive photon detection events [27]. The human ear, for example, is so sensitive that it

[8] Midget cells have a high spatial precision but a low temporal precision, while parasol cells have a low spatial precision but a high temporal precision.

operates at the limit imposed by Brownian movements of the stereocilia in the inner hear cells, which use mechanosensing to detect sound [69].

All these noise sources set high demands on neural information encoding in sensory organs. This is especially true, when the signal-to-noise ratio is small, for example in dark light conditions, low concentrations of odor molecules, or environments with high background noise. Shannon's mutual information automatically incorporates all noise effects in the noise entropy term (Eq. 2.6 in the previous section). Historically, the efficient coding hypothesis has been applied to populations of neurons, where the noise term has been ignored, resulting in ambiguous optimal solutions [23, 38, 41, 42]. The strength [40, 65, 71–75] and type [76, 77] of noise can have distinct effects on signal encoding. In the following section, I will review how noise affects parallel information encoding.

2.4.2 The influence of noise on parallel information encoding

The efficient coding hypothesis postulates that evolutionary pressure has led to the development of mechanisms that reduce noise and its deleterious effects on information encoding. One possibility to reduce the effect of noise is using parallel, redundant information channels: when the same signal is sent across N noisy channels and the output is summed, the signal strength adds up linearly with the number of channels N, while the noise strength increases only with $\sqrt{N}$ (since the noise is uncorrelated across the channels) [2]. Thus, the signal-to-noise ratio increases with $\sqrt{N}$. This is a comparatively flat increase, in particular when compared with the fact that in the case of no noise the information increases exponentially with N if channels are not used redundantly but optimally [25]. Nevertheless, it has been found consistently that redundant coding is optimal in the presence of high noise, as I will review now.

Several studies have used the efficient coding hypothesis to investigate how populations of neurons should encode a noisy signal. These studies used distinct encoding frameworks, noise models, neuron models, or constraints and often also optimized different functions (mutual information, Fisher information, stimulus reconstruction errors). Nevertheless, one common result of all of these studies is that in the regime of high noise it is optimal to have a redundant coding scheme, which "averages out noise" [29, 38, 42, 59, 62, 76, 78–86], while in the low noise regime it is optimal to either have independent coding [29, 59, 62, 76, 84–86] or decorrelation coding [38, 42, 79–83, 87]. This is in agreement with other studies that investigated filter sizes in single neurons and found large and monophasic filters – which means increased redundancy – to be optimal in the presence of high noise [40, 65, 72]. Note that the term "independent coding" comes from the fact that different parts of the signal – e.g. high vs. low stimulus values – are independently transmitted in different channels. This is contrasted to redundant coding, where the same signal is transmitted in all channels; and decorrelation coding where decorrelated parts of a correlated signal are separately transmitted in different channels. In the case of a correlated signal, decorrelation is the optimal strategy with low noise, while without a correlated signal no decorrelation is possible and therefore independent coding is the optimal strategy with low noise.

For populations of neurons, the difference between redundant and independent parallel coding can be interpreted as multiple neurons in the population that acquire the same response thresholds to average out uncertainties in stimulus representation due to noise – redundant coding – and neurons that acquire distinct thresholds to optimally encode different parts of the stimulus distribution – independent coding (Fig. 2.5A). This is in agreement with experimental data [29,84,86]. The intuitive explanation is that for low noise each neuron encodes a specific part of the stimulus space with high fidelity, thus giving high information about many parts of stimulus space. For high noise, on the other side, the information about the stimulus is heavily disturbed to such an extent that each single neuron cannot transmit much information anymore, and it makes more sense to pool neurons together to achieve at least some fidelity in *one* stimulus region.

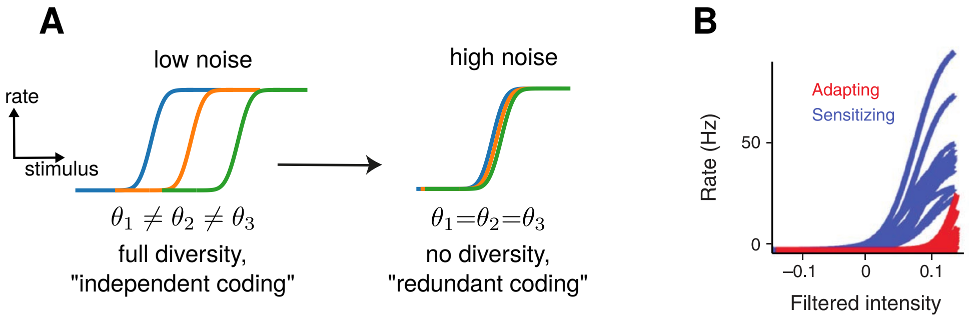


Figure 2.5: Efficient coding predicts diverse tuning curves at low noise levels and redundant tuning curves at high noise levels. A. Schematic tuning curves of three neurons show distinct thresholds θ at low noise for optimal information encoding (left). For high noise, the optimal solution is to have three equal thresholds (right). **B.** Measured tuning curves from retinal ganglion cells in salamander which respond to the same linearly filtered stimulus. Two distinct cell types are shown in color ("adapting", red; "sensitizing", blue) which systematically differ in threshold and maximum firing rate. From [29].

2.4.3 Convergence of parallel information streams

As I have reviewed in the previous section, encoding information with parallel streams using populations of neurons is advantageous compared to using a single stream. Another question when maximizing information between stimulus and output is how the output of the neuronal population converges further downstream. Both convergence (several channels merging into a smaller number of channels) and divergence (splitting of channels into a larger number of parallel channels) are seen across all stages of sensory processing [2]. Channel divergence means more neurons are needed, i.e. a higher need for space, material, and energy. Therefore, from an efficiency perspective, channel divergence is only favorable for high signal-to-noise ratios [2] and convergence is especially advantageous for low signal-to-noise ratios. This explains why there is much more convergence in the rod pathway compared to the cone[9] pathway in the retina [70].

Previous studies using the efficient coding hypothesis have assumed a framework in which the spiking output of a neural population converges, or is lumped, into a single output channel [85,86]. In contrast, other works have assumed a framework without signal lumping, i.e. where the signal is encoded by the independent spiking output of each neuron in the population [29,37,62,76]. In this thesis, I am going to compare these two commonly used frameworks for stimulus processing (Sec. 3.3).

2.5 Phase transitions indicate sudden qualitative changes of a system

For a population of two neurons, it has been found that the encoding scheme undergoes sudden changes when efficiently encoding a noisy stimulus [29]: the transition from independent coding at low noise to redundant coding at high noise happens not gradually but suddenly with increasing noise. These sudden changes show many similarities to phase transitions known from physics. Since I will also look into these sudden transitions in my thesis and compare them to phase transitions I will give the reader a brief introduction. If not stated otherwise, my sources of knowledge for this section are [89,90].

In physics, phases are different states of the same matter and a phase transition is a change between different states through the change of external variables like temperature or pressure. During a phase transition, the properties of the matter abruptly change. Well-known examples are the transition from ice to water or from water to vapor at the melting point or boiling point, respectively. Other examples are ferromagnets (above a critical temperature the magnetization suddenly disappears), superconductors (below a critical temperature they suddenly lose all electric resistance), superfluids (ultra-cold, liquefied Helium has zero viscosity[10]) and the miscibility of liquids (above a critical temperature two liquids become mixable). Different phases can be characterized through macroscopic variables; depending on the specific systems these are one of the following: density (in water-to-vapor

[9] Rods and cones are two types of photoreceptors, which transmit information about incoming light to bipolar cells. Rods are very light-sensitive and fast on the cost of a lower signal-to-noise ratio, while cones are less sensitive and slower but are less noisy. For example, each rod bipolar cell on average gets input from approximately 35 rods while each cone bipolar cell on average gets inputs from 2-6 cones [88].

[10] Viscosity quantifies the resistance of fluids to change their shape. Honey, for example, has a higher viscosity than water. Superfluid Helium shows such remarkable properties like zero viscosity – when being stirred it rotates indefinitely.

transitions), magnetization (ferromagnets), electrical conductibility (superconductors), viscosity (superfluids), or crystal structures (water-to-ice transition). The macroscopic variable of a system is called the *order parameter*. When an external parameter like temperature or pressure continuously changes, the properties of a system in general also change continuously. At a phase transition, however, the properties of the system abruptly change when the external parameter is just slightly varied. One of these properties is the order parameter, another is in general the symmetry of the system. The value of the external parameter at which phase transitions occur are called the *critical* value, e.g. *critical temperature* or *critical pressure*.

Phase transitions are extensively studied in thermodynamics and statistical physics. They are defined by a non-analytic behavior of the *free energy* of the system[11]. Non-analytic behavior means that the first or a higher-order derivative of the free energy with respect to temperature is discontinuous (i.e. jumps) or diverges (i.e. goes to infinity) (Fig. 2.6). Classically, phase transitions were classified according to Ehrenfest into two classes [91]: if the first derivative of the free energy is discontinuous the phase transition is of *first-order* (Fig. 2.6B), while if the second derivative is discontinuous the phase transition is of *second-order* (Fig. 2.6C). The modern classification groups all phase transitions that are not first-order into the class of *continuous* phase transitions and also includes divergent behavior of derivatives of the free energy (in contrast to just discontinuous behavior).

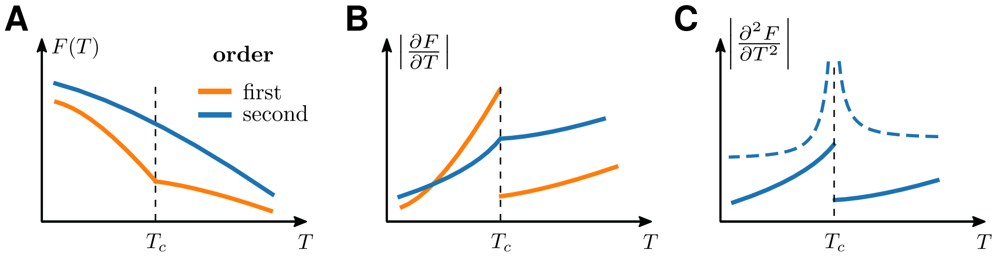


Figure 2.6: Schematic of free energy and its derivatives during phase transitions. A. Free energy F depending on temperature T for systems showing a first- or second-order phase transition (orange or blue color, respectively) at critical temperature T_c. **B.** First derivatives of the free energy curves shown in A. For the system showing the first-order phase transition, the first derivative of $F(T)$ is discontinuous at critical temperature. **C.** For the system showing the second-order phase transition, the second derivative is either discontinuous (solid) or diverges (dashed) at critical temperature.

In my studies of maximizing information transmission with populations of noisy

[11] The free energy is a quantity that all systems tend to minimize in the absence of outside force. If a system has minimized its free energy it is said to be in an *equilibrium state*. As a consequence, chemical reactions only happen spontaneously when the free energy is reduced during the reaction.

neurons, I discover behavior akin to phase transitions (Ch. 4). To find out how these phase transitions are related to phase transitions studied in statistical physics, I use measures like the *moment-generating function* of the neural responses, which is related to the moment-generating function of the energy. In the appendix A.1, I give a quick introduction to how phase transitions occur in statistical physics and how they are related to the moment-generating function of the energy.

3 Efficient coding in neural populations with multiple noise sources

Remark: Some of the methods, results, figures, and text in this chapter are part of an article entitled *Efficient population coding depends on stimulus convergence and source of noise* which has been written together with Shuai Shao and Julijana Gjorgjieva. The article has been uploaded to the preprint server *bioRxiv* [1] and is currently under review for publication in the journal *PLOS Computational Biology*. All methods, results, figures, tables, and text from that article which are part of this chapter were my contribution to the article. In this chapter, I specifically mention if and what I have taken from [1] before each (sub)section.

In the previous chapters, I presented the assumption that biological systems optimize their efficiency due to evolutionary pressure and that neural populations in sensory organs are optimized towards efficiently encoding information about the environment. Now, I will use that assumption to make a contribution in the understanding of why there is such a great variety of different neuron types in the sensory organs. I use a model that is based on a population of spiking neurons that encode a sensory stimulus under different noise scenarios. By maximizing its information encoding, I investigate how the individual neurons of the population optimally diversify in their nonlinear properties and how this diversification depends on the noise scenarios. Additionally, I will compare two different channel types that can be used for information encoding. These are, first, the *independent*-coding channel, where the full output of the neural population encodes the stimulus, and second, the *lumped*-coding channel, where the output is reduced to a scalar variable. Rather than investigating efficient coding in a specific sensory system, I sought to derive a general theoretical framework that applies to multiple sensory systems.

This chapter is organized as follows: first, I describe the framework of my model and carry out the mathematical calculations for the mutual information (Sec. 3.1), followed by a description of why and how I carry out numerically the calculation and maximization of the mutual information (Sec. 3.2). Afterward, I show how the maximized information depends on the noise and how and why the independent-coding channel outperforms the lumped-coding channel (Sec. 3.3). I also show how the individual nonlinearities of the neurons of the population have to be organized so that the information is maximized, including again a comparison between the two channel types (Sec. 3.4). Then, I will dedicate some attention to unexpected results, which seem faulty at first but still persisted after a battery of tests (Sec. 3.5). An investigation on how different noise models and different stimulus distributions affect the results is carried out in Sections 3.6 and 3.7, respectively. Finally, I summarize and briefly discuss my results (Sec. 3.8).

3.1 Framework and analytical expressions

Remark: The following framework was first published in [1] and content-wise is identical to [1]. This includes the methods (Eqs. 3.1-3.7), Fig 3.1, and the text. However, here I have slightly modified the text for stylistic reasons increased clarity.

A population of N neurons encodes a static, one-dimensional stimulus s drawn from a stimulus distribution $P(s)$ through the spike counts $\{k_1, ...k_N\} \equiv \vec{k}$ emitted in a coding time window ΔT (Fig. 3.1A). If not stated otherwise, I use a Gaussian stimulus distribution with mean zero: $s \sim \mathcal{N}(0, \sigma_s^2)$. The mapping from the stimulus value s to the spike count vector $\vec{k}$ happens through a set of N nonlinear functions (tuning curves) $\{\nu_1(s), ..., \nu_N(s)\}$, where $\nu_i(s)$ denotes the firing rate of the respective neuron i. The quality of the mapping from s to $\vec{k}$ is quantified by the mutual information $I_m(\vec{k}; s)$ between s and $\vec{k}$. Two noise sources, namely input noise (introduced at the input of the nonlinearities) and output noise (introduced at the output of the nonlinearity), distort the mapping and reduce the information. Both noise sources are described in detail below.

I assume that the nonlinearities take a filtered stimulus s as input. For example, if I assume that my sensory system of interest is the population of ANFs, which are highly nonlinear processing units [92], then s represents the stimulus value following preprocessing by the cochlea and the inner hair cells and not sound intensity when it reaches the eardrum. In particular, I assume that all linear filtering is included in this preprocessing.

I modeled the neurons' tuning curves as binary, each described by two firing rate levels $\{0, \nu_{\max}\}$ and an individual threshold θ_i which determines which of the two rates is the output of the tuning curve. Thus, the input-output functions of each neuron i can be represented by

$$\nu_i(x) = \nu_{\max}\Theta(\theta_i - x), \tag{3.1}$$

where $\Theta(.)$ is the *Heaviside function.*[1] The input to each nonlinearity is the sum of stimulus and input noise z: $x = s + z$. This simplification of a binary nonlinearity is justified by the fact that many sensory neurons have been described with steep tuning curves that resemble binary neurons [44, 59, 87], and it makes the problem mathematically traceable. Moreover, as I will explain below, input noise causes a

[1] The Heaviside function is a simple one-step function: $H(x) = 0$ for $x < 0$, and $H(x) = 1$ for $x \geq 0$.

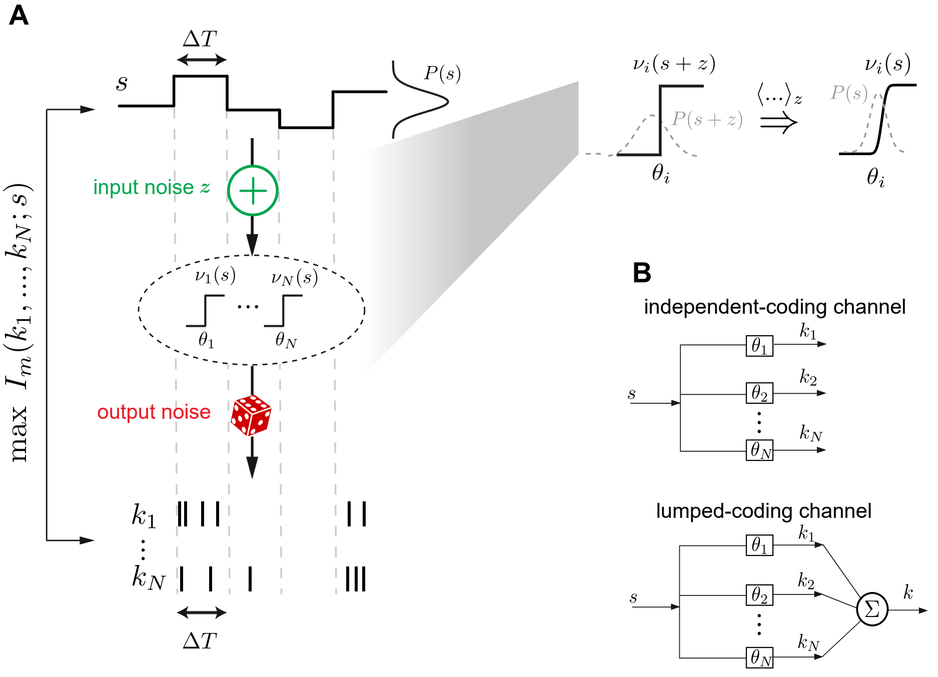


Figure 3.1: Stimulus encoding with a population of neurons in the presence of input and output noise. A. Framework: A static stimulus s (top) is encoded by a population of spike counts $\{k_1, ...k_N\}$ (bottom) in a coding time window ΔT. The stimulus is first corrupted by additive input noise z and then processed by a population of N binary nonlinearities $\{\nu_1, ...\nu_N\}$. Stochastic spike generation based on Poisson output noise corrupts the signal again. Thresholds $\{\theta_1, ..., \theta_N\}$ of the nonlinearities are optimized such that the mutual information $I_m(k_1, ..., k_N; s)$ between stimulus and spike counts is maximized. Inset: Introducing additive input noise and a binary nonlinearity can be interpreted as having a sigmoidal nonlinearity after the input noise is averaged, $\langle ... \rangle_z$. Shallower nonlinearities result from higher input noise levels. **B.** Two different scenarios of information transmission: With the independent-coding channel each neuron contributes with its spike count to the coding of the stimulus, while with the lumped-coding channel all spike counts are added into one scalar output variable that codes for the stimulus. From [1].

binary nonlinearity to appear as a sigmoidal. The question I mainly investigate in this thesis is about how the information encoding can be optimized, i.e. under which conditions the mutual information of the mapping from s to $\vec{k}$ is maximized. To answer this question, I focus on optimizing the nonlinearities in the framework, i.e. to find the optimal values of thresholds $\vec{\theta} \equiv \{\theta_1, ..., \theta_N\}$ in the population of N neurons:

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} I_m(\vec{k}; s). \tag{3.2}$$

As I will describe later, the optimal thresholds depend on the strength of the noise sources. Therefore, I will first introduce the two noise sources used in the framework.

Input noise

Before being processed by the nonlinearities, the stimulus s is corrupted by additive noise z drawn from a distribution $P(z)$. The size of input noise can be quantified by the ratio of its variance $\langle z^2 \rangle \equiv \sigma_z^2$ to the stimulus variance $\langle s^2 \rangle \equiv \sigma_s^2$. Without loss of generality, I set $\sigma_s^2 = 1$ and thus $\sigma^2 := \sigma_z^2/\sigma_s^2$ alone stands for the size of input noise. The noise affects the stimulus independently for each nonlinearity, i.e. I did not consider correlated noise since previous work has shown that the case of correlated noise can be reduced to independent noise with lower σ^2 [76]. Similarly to the stimulus distribution, I primarily examined the case with the noise drawn from a Gaussian distribution, $z \sim \mathcal{N}(0, \sigma^2)$, but I also considered other distributions with different kurtosis $\langle z^4 \rangle$ (see Sec. 3.6.1). Since the input to the nonlinearities is $x = s + z$, the effective tuning curves, $\nu_i(s)$, can be described to have a sigmoidal shape (Fig. 3.1A, inset). Note that here the threshold θ is not the input value where the effective tuning curve indicates a firing rate significantly different from zero, but the input value where the tuning curve has the steepest part, i.e. at the inflection point, where $\nu = \nu_{\max}/2$. A larger input noise size, determined by the variance of the noise σ^2, corresponds to a shallower slope of the tuning curve, without effecting the threshold value. In the remainder of the text, I use the standard deviation σ to refer to the size of input noise.

Output noise

Output noise was implemented by generating output spikes stochastically. In general, I use Poisson output noise, for which each of the spikes counts k_i in a coding window ΔT given firing rate ν_i is Poisson distributed, i.e.

$$P(k_i|\nu_i) = \frac{(\nu_i \Delta T)^{k_i}}{k_i! \; e^{-\nu_i \Delta T}}. \tag{3.3}$$

Output noise models different from Poisson noise are discussed in Sec. 3.6.4. Here, large output noise corresponds to the case when the product of $\nu_{\max}$ and ΔT is small; in this case the output of a given cell i is often $k_i = 0$ making it more difficult to distinguish whether the underlying firing rate for that cell is 0 and thus the stimulus was smaller than the threshold θ_i, or whether the firing rate is $\nu_{\max}$ and the stimulus was greater than θ_i. The output noise size can thus be quantified by the expected[2] spike count for maximum firing rate, $R := \nu_{\max}\Delta T$, where small R means high noise since for small R there is a higher ambiguity about the real firing rate. For $R \to 0$ the two firing rate levels 0 and $\nu_{\max}$ are indistinguishable, meaning all information about the stimulus is lost (case of infinite output noise).

[2] For the Poisson distribution, the mean is given as the rate parameter, which here is $\nu_i \Delta T$.

The implementation of output noise described in this section can be understood as a constraint on the maximum firing rate level $\nu_{\max}$ while having a fixed coding window length ΔT.

Maximizing mutual information for two different coding scenarios

In addition to two noise sources, I further distinguish between two different scenarios previously considered in the literature for how the sensory signal converges after being processed by a population of neurons. First, the scenario described so far, termed the *independent-coding channel*, where a vector of spike counts $\vec{k} = \{k_i\}$ generates a population code of the stimulus (Fig. 3.1B, top). Here, each spike count independently contributes to the total information [29,62,76]. Second, the *lumped-coding channel* where a scalar output variable $k = \sum_i k_i$, obtained by summing the individual spike counts k_i, codes for the stimulus [85,86] (Fig. 3.1B, bottom). For given noise levels σ and R the goal is to find nonlinearities that optimally encode the stimulus s with a vector of spike counts $\vec{k} \equiv \{k_i\}$ (independent-coding channel) or the lumped spike count $k = \sum k_i$ (lumped-coding channel). As a measure for optimality for the independent- and lumped-coding channels I choose the mutual information between stimulus s and observed spike count $\vec{k}$ or k, respectively. In other words, the optimization of the threshold vector $\vec{\theta}$ stated in Eq. 3.2 is performed for both channel types.

The mutual information gives an upper bound on how much information can on average be obtained about the input by observing the output. As described in Section 2.3, it is given as the difference between output entropy $H(\vec{k})$ and noise entropy $H(\vec{k}|s)$ [25]:

$$
\begin{aligned}
I_m(\vec{k};s) &= H(\vec{k}) - H(\vec{k}|s) &(3.4)\\
&= -\sum_{k_1=0}^{\infty} \ldots \sum_{k_N=0}^{\infty} P(\vec{k}) \log_2\left(P(\vec{k})\right) + \sum_{k_1=0}^{\infty} \ldots \sum_{k_N=0}^{\infty} \int_s \mathrm{d}s\, P(s)\, P(\vec{k}|s) \log_2\left(P(\vec{k}|s)\right) \\
&= \sum_{k_1=0}^{\infty} \ldots \sum_{k_N=0}^{\infty} \int_s \mathrm{d}s\, P(s)\, P(\vec{k}|s) \log_2\left(\frac{P(\vec{k}|s)}{\int_{s'} \mathrm{d}s' P(s') P(\vec{k}|s')}\right) &(3.5)
\end{aligned}
$$

where the input-output kernel $P(\vec{k}|s)$ is the probability of obtaining a certain vector of output spikes for a given stimulus value. In the case of the lumped-coding channel, the calculations are the same, except that the spike count is now one-dimensional, i.e. I have $I_m(k;s)$ as the mutual information and $P(k|s)$ as the input-output kernel:

$$
\begin{aligned}
I_m(k;s) &= H(k) - H(k|s) &(3.6)\\
&= \sum_{k=0}^{\infty} \int_s \mathrm{d}s\, P(s)\, P(k|s) \log_2\left(\frac{P(k|s)}{\int_{s'} \mathrm{d}s' P(s') P(k|s')}\right) &(3.7)
\end{aligned}
$$

The way of calculating the respective kernels differ between the channel types and are now derived in detail.

3.1.1 Mutual information of the independent-coding channel

Remark: The analytical expressions, methods, and text in this subsection (in particular, Eqs. 3.8-3.21) were originally developed in [1] and content-wise are identical to [1]. However, here I have slightly modified some text for stylistic reasons and increased clarity.

In the case of the independent-coding channel, information about the input stimulus s is encoded by the N-dimensional spike count vector $\vec{k}$. The input-output kernel $P(\vec{k}|s) \equiv P(k_1, ..., k_N|s)$ can be expressed by multiplying $P(k_1, ..., k_N|\nu_1, ..., \nu_N)$ and $P(\nu_1, ..., \nu_N|s)$ and summing over all possible combinations of firing rate states $\{0, \nu_{\max}\}^N$:

$$P(\vec{k}|s) = \sum_{\nu_1 \in \{0,\nu_{\max}\}} ... \sum_{\nu_N \in \{0,\nu_{\max}\}} P(k_1, ..., k_N|\nu_1, ..., \nu_N) P(\nu_1, ..., \nu_N|s) \tag{3.8}$$

I assume no noise correlations and thus ν_i conditional on s are independent of each other:

$$P(\nu_1, ..., \nu_N|s) = \prod_i P(\nu_i|s) \tag{3.9}$$

Furthermore, all k_i are independent of each other conditional on a set of firing rates $\{\nu_1, ..., \nu_N\}$, and every k_i only depends on $\nu_{j=i}$:

$$P(k_1, ..., k_N|\nu_1, ..., \nu_N) = \prod_i P(k_i|\nu_1, ..., \nu_N) = \prod_i P(k_i|\nu_i) \tag{3.10}$$

Taken together:

$$P(\vec{k}|s) = \sum_{\vec{\nu} \in \{0,\nu_{\max}\}^N} \prod_i^N P(k_i|\nu_i) P(\nu_i|s) = \prod_i^N \sum_{\nu_i \in \{0,\nu_{\max}\}} P(k_i|\nu_i) P(\nu_i|s) \tag{3.11}$$

$P(k_i|\nu_i)$ follows a Poisson distribution (Eq. 3.3) and $P(\nu_i|s)$ denotes the probability of having a firing rate of zero (or $\nu_{\max}$) for a given stimulus s. Since the input noise fluctuations are on a much faster time scale than the length of the coding window (over which the stimulus is assumed to be constant), an averaging over z can be performed. Thus $P(\nu_i = 0|s)$ (or $P(\nu_i = \nu_{\max}|s)$) is given as the probability that stimulus plus noise is smaller (or larger, respectively) than threshold θ_i, which is the area under the noise distribution for which $s + z < \theta_i$ (or $s + z \geq \theta_i$, respectively):

$$P(\nu_i = \nu_{\max}|s) = \int_{\theta_i - s}^{\infty} \mathrm{d}z P_z(z) =: H_i(s), \tag{3.12}$$

$$P(\nu_i = 0|s) = \int_{-\infty}^{\theta_i - s} \mathrm{d}z P_z(z) = 1 - H_i(s). \tag{3.13}$$

$H_i(s)$ can be viewed as the "effective" tuning curve that one would measure electrophysiologically (see also Fig. 3.1, top right). It is the cumulative distribution function of the dichotomized noise distribution. If the noise distribution is normal distributed with variance σ^2, the effective tuning curve is given by the complementary error function[3]:

$$H_i(s) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\theta_i - s}{\sqrt{2}\sigma}\right). \tag{3.14}$$

Then one can calculate the mutual information by performing the summation over all output variables $k_1, ..., k_N$. The output noise is included since $P(k_i|\nu_i)$ is Poisson distributed. According to the Poisson distribution, $P(k_i = 0|\nu_i = \nu_{\max}) = e^{-R}$. For each neuron i, all spike counts greater than zero can be lumped into one state due to the fact that if there are one or more spikes emitted, the firing rate ν_i cannot be zero for any $k_i > 0$ but must be $\nu_{\max}$ since $P(k_i > 0|\nu_i > 0) = 0$ for $\nu_i = \{0, \nu_{\max}\}$. This state is denoted as 1 and from now on I have $k_i \in \{0, 1\}$. Thus

$$P(k_i = 1|\nu_i = \nu_{\max}) = 1 - P(k_i = 0|\nu_i = \nu_{\max}) = 1 - e^{-R}. \tag{3.15}$$

The mutual information can then be calculated as

$$I(\vec{k}; s) = \sum_{k_1,...,k_N \in \{0,1\}^N} \int_s P_s(s) \prod_{i=1}^{N} \sum_{\nu_i \in \{0,\nu_{\max}\}} P(k_i|\nu_i)P(\nu_i|s)\, \log_2\left(\frac{...}{\int_{s'} \mathrm{d}s' P(s')...}\right) \mathrm{d}s \tag{3.16}$$

where the placeholder "..." indicate that $P(k_i|s)$ shall have the same expression as outside the logarithm, namely:

$$P(k_i|s) = \prod_{i=1}^{N} \sum_{\nu_i \in \{0,\nu_{\max}\}} P(k_i|\nu_i)P(\nu_i|s)\ , \tag{3.17}$$

and with

$$\sum_{\nu_i} P(k_i = 0|\nu_i)P(\nu_i|s) = (1 - H_i(s)) + \mathrm{e}^{-R} H_i(s) =: Q_i(s), \tag{3.18}$$

$$\sum_{\nu_i} P(k_i = 1|\nu_i)P(\nu_i|s) = (1 - \mathrm{e}^{-R}) H_i(s) =: S_i(s) \tag{3.19}$$

where output noise is $R = \nu_{\max}\Delta T$ as defined earlier. Taken together, the mutual information for the independent-coding channel is

$$I(\vec{k}; s) = \sum_{k_1=0}^{1} ... \sum_{k_N=0}^{1} \int_s P_s(s) \left(\prod_{i=1}^{N} P_{k_i}(s)\right) \log_2\left(\frac{\prod_i P_{k_i}(s)}{\int_{s'} \mathrm{d}s' P(s') \prod_i P_{k_i}(s')}\right) \mathrm{d}s \tag{3.20}$$

with

$$P_{k_i}(s) = \begin{cases} Q_i(s), & \text{for } k_i = 0 \\ S_i(s), & \text{for } k_i = 1 \end{cases} \tag{3.21}$$

[3] The error function, $\mathrm{erf}(x)$, is defined as the integral from $-\infty$ to x over a Gaussian distribution, while the complementary error function, $\mathrm{erfc}(x)$, is defined as the integral from x to ∞ over a Gaussian distribution.

3.1.2 Mutual information of the lumped-coding channel

Remark: The analytical expressions, methods, and text in this subsection (in particular, Eqs. 3.23-3.29) were originally developed in [1] and content-wise are identical to [1]. However, here I have slightly modified some text for stylistic reasons and increased clarity.

In the case of the lumped-coding channel, the output spike counts of all cells are lumped together and the information about the input stimulus is encoded by the one-dimensional variable $k = \sum_i k_i$ (Fig. 3.1B). For the case of only input noise, where $Q_i(s) = 1 - H_i(s)$ (Eq. 3.18) and $S_i(s) = H_i(s)$ (Eq. 3.19), previous work took advantage of the fact that only two output states exist for each neuron and explained how $P(k|s)$ can be calculated using a recursive formula [85, 93]. I have extended these calculations to additional Poisson output noise, where all neurons can have potentially infinite output states. I write $P(k|s)$ as $P(k|N,s)$ and use the notation by McDonnell et al. [85, 93], who define the probability of having k spikes with N neurons and stimulus value s as

$$T^N_{k,s} := P(k|N,s). \tag{3.22}$$

Furthermore, they define $P_{k_i|s,i}$ as the probability of cell i firing k_i spikes in a coding window ΔT when the stimulus is s. Using this basis, I can now add output noise by expressing the probability of having k spikes with N cells as the probability of having k_N spikes by the N-th neuron multiplied by the probability of having $k - k_N$ spikes by the other neurons and taking into account all possibilities of k_N by summing over k_N:

$$T^N_{k,s} = \sum_{k_N=0}^{k} P_{k_N|s,i=N} \cdot T^{N-1}_{k-k_N,s} \tag{3.23}$$

where

$$\begin{aligned} P_{k_i|s,i} &= \sum_{\nu_i \in \{0,\nu_{\max}\}} P(k_i|\nu_i)P(\nu_i|s) && (3.24)\\ &= P(k_i|\nu_i = 0)P(\nu_i = 0|s) + P(k_i|\nu_i = \nu_{\max})P(\nu_i = \nu_{\max}|s) && (3.25)\\ &= \begin{cases} (1 - H_i(s)) + e^{-R}H_i(s), & \text{for } k_i = 0 \\ \frac{(R)^{k_i}}{k_i!\, e^{-R}} H_i(s), & \text{for } k_i > 0 \end{cases} && (3.26) \end{aligned}$$

is the probability of cell i emitting k_i spikes given stimulus s, and

$$T^N_{0,s} = \prod_{i=1}^{N} P_{0|s,i} = (1 - H_i(s)) + \mathrm{e}^{-R}H_i(s) = \prod_{i=1}^{N} Q_i(s) \tag{3.27}$$

being the probability of having zero spikes with N cells, as well as

$$T^1_{k,s} = P_{k_1|s,i=1} \tag{3.28}$$

being the probability of having k spikes with $N = 1$. With this recursive procedure, the expression for the mutual information of the lumped channel

$$I_m(k;s) = \sum_{k=0}^{\infty} \int_s \mathrm{d}s\, P(s)\, T_{k,s}^N \log_2 \left(\frac{T_{k,s}^N}{\int_{s'} \mathrm{d}s' P(s') T_{k,s'}^N} \right) \tag{3.29}$$

can be calculated. For every $k = 0, 1, 2, ...$ until an upper bound which is determined by the precision one wants to reach,[4] $T_{k,s}^N$ is calculated for every $k_N = 0, 1, ...k$ by using the recursive formula in Eq. 3.23. This is computationally very expensive for larger R (due to non-vanishing contributions even for very large k) and larger N (since all combinations of spike outputs that lead to a single k have to be considered). Thus, I studied only populations with up to $N = 3$ neurons and with R – the expected spike count in ΔT in the case of firing rate $\nu_{\max}$ – up to[5] $R = 10$. As with the independent-coding channel, input noise σ is included in $H_i(s)$ (see Eq. 3.12) and the output noise level is denoted by R.

3.2 Numerical calculations of the mutual information

The goal is to find the optimal thresholds $\vec{\theta}$ that maximize mutual information for given levels of input and output noise σ and R:

$$\vec{\theta}^* = \underset{\vec{\theta}}{\arg\max}\; I_m(k; s|\sigma, R). \tag{3.30}$$

Since the expressions for the mutual information here (Eqs. 3.20 and 3.29) contain integrals over error functions or recursive formula, they have to be computed numerically. As a consequence also the optimization has to be performed numerically.

3.2.1 Numerical integration

As can be seen from Eqs. 3.20 and 3.29, numerical integration has to be performed twice for each evaluation of mutual information. These numerical integrations happen to be the most computationally expensive parts. Since during the numerical optimization a high number of evaluations of the mutual information are necessary, it was important to make the integration process as quick as possible. After testing several different algorithms (Riemann, Trapezoid, Romberg, Simpson, and adaptive algorithms; [94]) for speed, I settled for the Trapezoid algorithm. In Sec. 3.5.2, I will show that errors arising from numerical integration are negligible. To avoid numerical instabilities in the case that $\lim_{x\to 0} x \log(x)$, I always add a machine epsilon, $\epsilon \approx 2.2 \cdot 10^{-16}$ [95], to $P(k|s)$, and $P(\vec{k}|s)$, respectively.

[4] I usually stop increasing k when contributions become smaller than 10^{-12} bits

[5] Note that calculating just *one* $P(k|s)$ for $N = 3$ and $R = 10$ requires on the order of 50 000 evaluations of Eq. 3.24. To compute the information for *one* combination of thresholds, this has to be done $k(k-1)/2$ times where k is in the order of approximately 80. That means around 150 million evaluations of Eq. 3.24, which again has to be performed many times during optimization and for 200 different values of σ.

3.2.2 Numerical optimization

As for numerical integration, several algorithms exist for numerical optimization. However, optimization seems to be more involved and thus a variety of different algorithms exists, often specialized in specific properties of the function to optimize [96]. After systematically exploring several of them, I chose the Nelder-Mead simplex (NM) algorithm and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, both implemented in the Scipy package [97, 98]. Both of them are local optimizers – which means they, in general, find only the (local) maximum close to the initial point.[6] I tried global optimizers like basin-hopping but were found to be too slow – at least without extensive adjustments of optimizer parameters.

The NM algorithm does not rely on estimating the gradient of the objective function, while the BFGS algorithm relies on calculating or estimating the inverse of the Hessian matrix (see Eq. 4.3 in Sec. 4.1) to estimate the gradient. I found the BFGS algorithm, in general, to be faster than the NM algorithm, however, for the independent-coding channel, the BFGS algorithm becomes problematic around critical noise values where one eigenvalue of the Hessian approaches zero (see Sec. 4.3) and thus leads to large numerical imprecisions when inverting the Hessian. For the lumped-coding channel this is unproblematic (see Sec. 3.4.5 and Sec. 4.3). Thus for speed purposes, I used an adaptation of the BFGS algorithm implemented in Scipy [98] for the lumped-coding channel and an adaptation of the NM algorithm for the independent-coding channel [97]. In Sections 3.4.5 and 3.5.2, I show that in general for my calculations the specific optimizer is not very important – except for speed purposes – as most of them lead to basically identical results and local maxima can easily be spotted.

To spot possible local maxima (which are somewhat prevalent for $N \geq 4$), I applied a grid of initial values. After some trials, it was possible to predict what form of initial values lead to local maxima. Furthermore, in general, local maxima can be easily spotted and checked by looking at the plots showing optimal thresholds [1]. I will describe this prevention of local maxima in more detail not here but in Section 3.4.4 after the reader has become familiar how to read the plots showing optimal thresholds. If not mentioned specifically otherwise, all results in my work show global maxima.

The heavy numerical procedure limited my analysis to small population sizes with a maximum of three neurons in the case of the lumped-coding channel and six neurons in the case of the independent-coding channel [1].

[6] More precisely, they find the local maximum in whose basin the initial point lies.

3.3 Maximal mutual information of the respective channel types

Remark: The results of this section (Fig. 3.2 and the text) have been published before in [1] and content-wise are identical to [1]. However, here I have slightly modified some text for stylistic reasons and increased clarity.

In this section, I will show how maximized information values for the lumped- and independent-coding channels for a population of $N = 3$ neurons depend on both input and output noise values. Then I will compare the performance of the two channel types with each other.

As expected, the smaller the noises, the higher is the maximized mutual information; both for the independent-coding and the lumped-coding channel (Fig. 3.2A and B, respectively). Additionally, I found that the increase of information with smaller output noise (higher R) saturates faster in the case of the independent-coding channel, as can be seen by the flattening of the contour lines as R increases (compare Fig. 3.2A and B). Note that the optimal thresholds that lead to these values of maximal mutual information are shown in Figure 3.4 and described in Section 3.4. I quantified the ratio and the absolute differences in information transmission between the two channels (Fig. 3.2C,D). For all finite input and output noise levels, the independent-coding channel outperforms the lumped-coding channel. The information loss due to lumping is the largest at intermediate levels of output noise and low levels of input noise; for instance, at $R \approx 2.5$ and $\sigma \approx 0$ the independent-coding channel transmits up to 40% more information than the lumped-coding channel (Fig. 3.2D). To best visualize these differences, I fixed one source of noise and varied the other. For fixed output noise R, the information loss in the lumped-coding channel relative to the independent-coding channel monotonically decreases as a function of the input noise σ (Fig. 3.2E). There, I also included the case of zero output noise ($R \to \infty$), showing that in this special case the two channels transmit basically the same amount of information.[7] The difference in information transmitted by the independent- and lumped-coding channels as a function of the output noise R for fixed input noise σ demonstrates that the information loss due to lumping is a non-monotonic function of output noise R (Fig. 3.2F), with the largest loss occurring in the biologically realistic range of intermediate noise [87, 99]. In summary, I found that in the presence of both input and output noise, the lumped-coding channel transmits less information than the independent-coding channel, especially for intermediate output noise values. In the following subsections, I will give an intuitive explanation of why this is the case and what it potentially implicates for sensory systems in biology.

[7] For the lumped-coding channel it is computationally very expensive to have large values of R, but that $R \to \infty$ is computationally cheap.

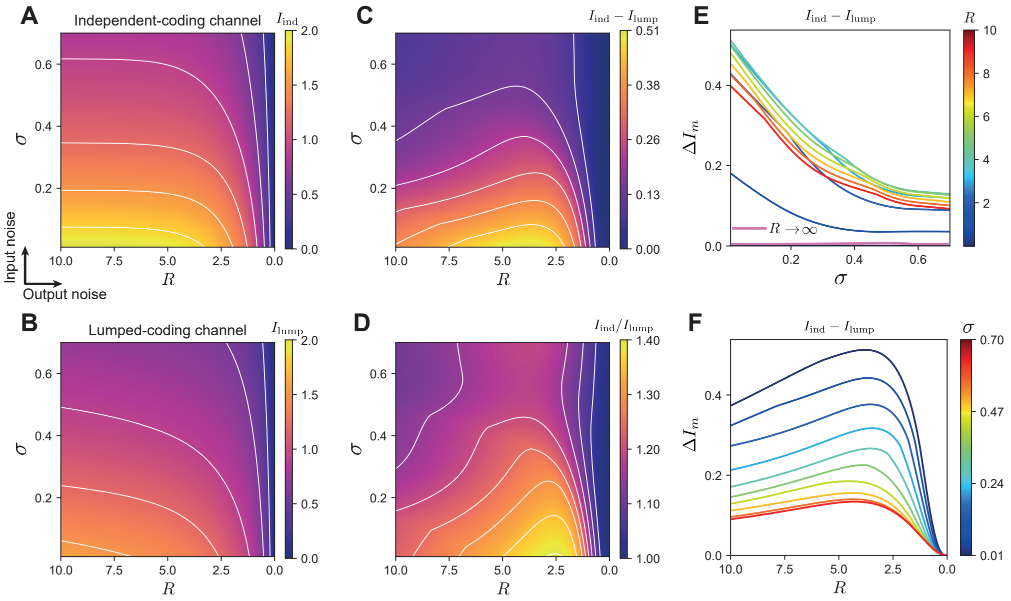
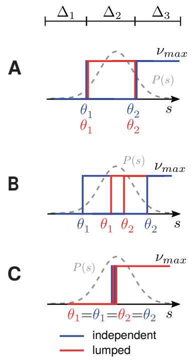


Figure 3.2: Maximized mutual information for the lumped- and independent-coding channels for a population of three neurons. All information units are in bits. **A.** Information of the independent-coding channel is color-coded for different combinations of output noise R and input noise σ. White contours indicate constant information. **B.** Information of the lumped-coding channel. **C.** Absolute information difference between the two coding channels. **D.** Information ratio between the two coding channels. Both C and D show a region of intermediate output noise where the independent-coding channel substantially outperforms the lumped-coding channel. **E.** Information difference depending on input noise σ for various levels of output noise R, corresponding to vertical slices from C. In addition, the case for no output noise ($R \to \infty$) is shown for comparison. **F.** Information difference depending on output noise R for various levels of input noise σ, corresponding to horizontal slices from C. Adapted from [1].

3.3.1 Why the independent channel encodes more information

Remark: The results of this section (Fig. 3.3 and the text) have been originally published in [1] and content-wise are identical to [1]. However, here I have slightly modified some text for stylistic reasons and increased clarity.

To gain intuition on why lumping causes an information loss and why this loss is highest for intermediate output noise while it vanishes in the limits of zero and infinite output noise, I illustrate the case with vanishing input noise ($\sigma = 0$) and a population with two neurons with thresholds $\theta_1 < \theta_2$, which divide the entire stimulus distribution into three regions: $\Delta_1 : s < \theta_1$, $\Delta_2 : \theta_1 \leq s < \theta_2$ and $\Delta_3 : s \geq \theta_2$ (Fig. 3.3, left). Here, I compute all possible spike counts and corresponding "estimation probabilities," $P(s \in \Delta_i | \vec{k})$ which describe the probability of the stimulus being in each of the three regions $\{\Delta_i\}_{i=\{1,2,3\}}$ for a given spike count $\vec{k}$ (Fig. 3.3).



		Spike count			Estimation probab. $P(s \in \Delta_j \mid k, \vec{k})$		
		k	k_1	k_2	$P(s \in \Delta_1)$	$P(s \in \Delta_2)$	$P(s \in \Delta_3)$
vanishing output noise $(R \to \infty)$	independent and lumped	0	0	0	1	0	0
		R	R	0	0	1	0
		$2R$	R	R	0	0	1
intermediate output noise assume $k = 1$	independent	1	1	0	0	α	$1-\alpha$
		1	0	1	0	0	1
	lumped	1	?	?	0	α'	$1-\alpha'$
high output noise $(R \to 0)$	independent and lumped	0	0	0	β	-	$1-\beta$
		1	1	0	0	-	1
		1	0	1	0	-	1

Figure 3.3: Schematic illustrating the dominance in information of the independent- over the lumped-coding channel. Here, I treat the case of $N = 2$ cells and vanishing input noise ($\sigma = 0$). Left schematics: The relative positions of optimal thresholds of both the independent- and lumped-coding channel are shown in red and blue, respectively (see also Fig. 3.4). **A,C.** For the limits of vanishing and very high output noise, the stimulus "estimation probabilities" $P(s \in \Delta_j|\vec{k})$ are identical. **B.** In contrast, in the case of intermediate noise levels these estimation probabilities are different, as illustrated when the total number of spikes $k = k_1 + k_2 = 1$. Yellow shading shows where the noise entropy is higher in the lumped-coding channel. α, α', and β denote probability values which depend on the exact noise level. Red and blue color indicate values for lumped- and independent-coding channel, respectively. From [1].

These estimation probabilities vary as a function of output noise, and I consider three cases: negligible, intermediate, and very high output noise. First, in the limit of vanishing output noise where $R = \nu_{\max}\Delta T \to \infty$ the information encoded by both channels is identical because with optimal thresholds both reach capacity and transmit $\log_2(N + 1 = 3) \approx 1.6$ bits of information (Fig. 3.3A). In particular, whenever the stimulus is larger than the threshold of a given cell, that cell will on average fire R spikes. Since $R \to \infty$, for that given cell the probability of having 0 spikes is infinitesimal. This unambiguously determines the stimulus region $\{\Delta_i\}$ in which the stimulus occurs. Hence, the estimation probabilities all become either 0 or 1, leading to identical output entropy for both coding channels, and consequently identical mutual information with zero noise entropy.

For intermediate output noise, the independent- and the lumped-coding channels have distinct estimation probabilities. Although in principle the number of emitted spikes can be anything, let me consider the example where the total number of spikes is 1 ($k_1 + k_2 = 1$, Fig. 3.3B). I demonstrate that the lumped-coding channel loses information because knowledge about the identity of which individual cell spiked is lost. For example, if the cell with higher threshold θ_2 fires at least one spike, this implies with certainty that the stimulus is greater than θ_2. The lumped-coding channel fails to encode this information since in principle the spike could also have been emitted by the cell with lower threshold θ_1. Thus, the estimation probabilities α' and $1-\alpha'$ for the stimulus being below or above θ_2, respectively, are nonzero. For the independent-coding channel, however, the corresponding estimation probabilities α and $1-\alpha$ are nonzero and non-one only if the cell with the lower threshold θ_1 fires a spike, but not if the cell with the higher threshold θ_2 fires a spike. Therefore, for the independent-coding channel, there are more cases in which the uncertainty is resolved, leading to higher mutual information. As an example, for output noise of $R = 2.5$, the mutual information for the independent- and lumped-coding channel is 1.30 and 1.01 bits, respectively (for optimized thresholds).

For very high output noise, $R \to 0$, the expected spike count of either of the cells is very small, even when the stimulus is larger than the respective threshold with the resulting firing rate ν_{max} (Fig. 3.3C). This means that most of the time the observed spike count of each cell is 0, rarely 1, and never 2 (the probability of observing more than one spike is infinitesimal). In this high noise regime, the optimal solution for both the independent- and lumped-coding channels is to make both thresholds identical, i.e. $\theta_1 = \theta_2$ ("redundant coding", see Sec. 2.4.2 and the results of optimal thresholds in the next section). Therefore, the intermediate regime Δ_2 does not exist for very high output noise and the two possibilities of having a spike from either cell are equivalent. Thus, if the observed spike count is 1, then there is no possibility of error for either channel. Similarly, if the observed spike count is 0, the two different estimation probabilities are the same for both channels, namely β and $1-\beta$ for the stimulus being above or below $\theta_{1,2}$, respectively. This results in identical mutual information between stimulus and response for both channels.

In summary, one can say that the contribution of each neuron to the overall spike count provides additional information about the stimulus, which is lost by summing all the spike counts through lumping. Thus, lumping acts as another form of noise, since it corrupts information transmission. Now I will give reasons why it could still be beneficial for biological systems to use lumped-coding channels.

3.4 Optimal thresholds which maximize information encoding

In this section, I will present the optimal population thresholds for which the spiking output of the populations achieves maximal information about the stimulus. I first discuss both channel types with three neurons, for which the maximized information has already been shown in the previous section. Afterwards, I will present optimal thresholds when I extend the number of neurons for up to six for the independent-coding channel. Furthermore, I will show that in the limit cases of one noise value going to zero my results are equal to results from the literature. Then, I will describe how in the optimizing process I avoided mistaking suboptimal thresholds of local maxima of the information with truly optimal thresholds of the global maximum. Finally, I will show for the lumped-coding channel that a gradient-based optimization algorithm provides the same results as an optimizer based on the simplex method.

3.4.1 General results for lumped- and independent-coding channel

Remark: The results of this subsection (Fig. 3.4 and the text) have been originally published in [1] and content-wise are identical to [1]. However, here I have slightly modified some text for stylistic reasons and increased clarity.

For both the independent- and lumped-coding channel with three neurons, the optimal number of distinct thresholds in the population depends on the source and level of noise (Fig. 3.4). When both sources of noise are negligible, the optimal number of thresholds is three, representing a fully diverse population where all thresholds are distinct. However, when both input and output noise is high, the optimal number of thresholds in the population is one, representing a fully redundant population where all thresholds are equal. This is true for both the independent- and lumped-coding channel (Fig. 3.4A and F, respectively) as it is in accordance with many previous studies (see Sec. 2.4.2). The most interesting cases arise at intermediate input and output noise levels, where I found two distinct optimal thresholds, i.e. two thresholds are equal and the third threshold has a different value.

To gain a better understanding of the transition between different threshold regimes as a function of noise, I fix one level of noise and examine the thresholds as a function of the other noise level. I found that the number of distinct thresholds in the population generally decreases with increasing input or output noise through a set of bifurcations. I call the noise levels at which these bifurcations in the thresholds appear *critical* noise levels. For the lumped-coding channel, the threshold bifurcations occur at lower noise levels compared to the independent-coding channel, i.e. the critical noise levels of the lumped coding channel are smaller than those of the independent channel. This result makes intuitive sense because as was pointed out in the previous section (Sec. 3.3.1), lumping multiple information pathways into a single coding channel reduces the possible values of the encoding variable and increases the noise entropy, and therefore acts like an additional noise source. Thus the effective critical noise levels – by taking into account the noise-like effect of lumping – of the lumped- and independent-coding channel might actually be in a similar range.

Optimal thresholds of the independent-coding channel

For the independent-coding channel, the thresholds become distinct from each other gradually, in the sense that the differences between the optimal thresholds change continuously, both as a function of output noise when the input noise level is fixed (Fig. 3.4B) and also as a function of input noise when the output noise level is fixed

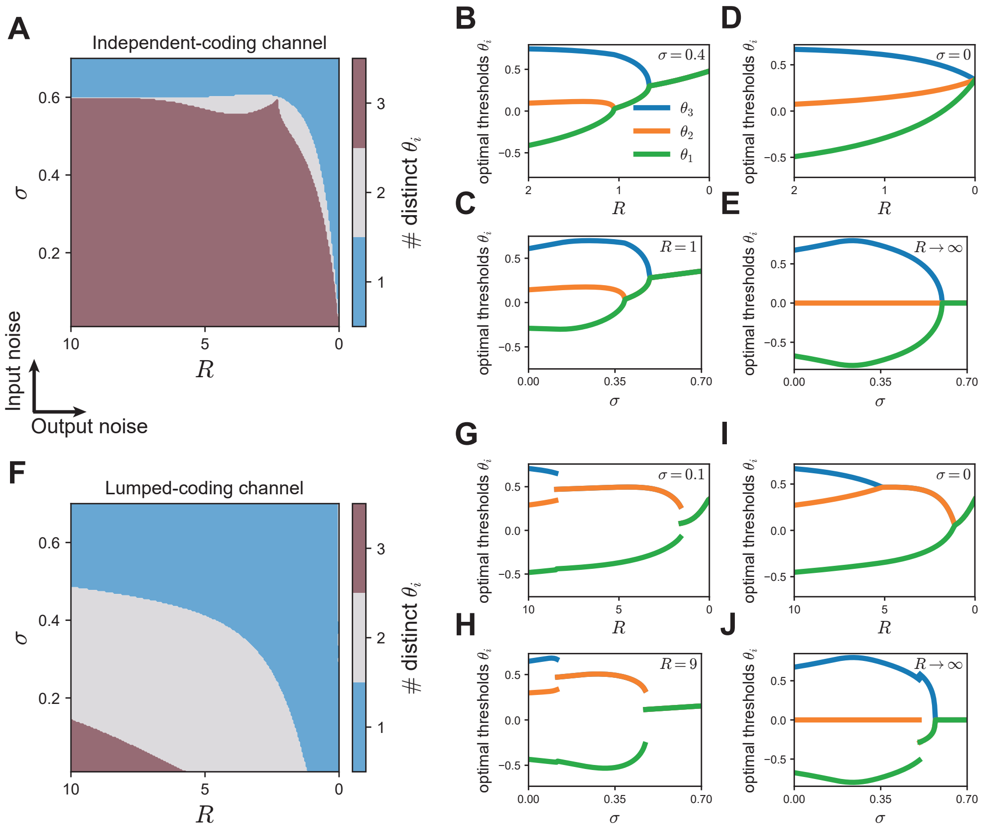


Figure 3.4: Optimal thresholds for the independent- and lumped-coding channels. Optimal thresholds for the independent-coding channel (A-E) are compared to the lumped-coding channel (F-J) for a population of $N = 3$ neurons. **A.** The optimal number of distinct thresholds depend on input noise σ and output noise R. **B.** The optimal thresholds as a function of output noise for a fixed value of input noise ($\sigma = 0.4$). **C.** The optimal thresholds as a function of input noise for a fixed value of output noise ($R = 1$). **D.** The optimal thresholds as a function of output noise in the limit of no input noise ($\sigma = 0$). **E.** The optimal thresholds as a function of input noise in the limit of vanishing output noise ($R \to \infty$). **F-J.** As (A-E) but for the lumped-coding channel. Intermediate noise levels in (G,H) take smaller values of R and σ in the lumped-coding channel since lumping itself acts like a source of noise (G: $\sigma = 0.1$, H: $R = 9$). Note the different scaling of the R-axis in (G,I) compared to (B,D) and the different values of the fixed noise source in (B,C) compared to (G,H). From [1].

(Fig. 3.4C). In the case when one source of noise is zero, these bifurcations represent the transition from all optimal thresholds being distinct directly to the state where all optimal thresholds are identical, without an intermediate state where two thresholds are the same (Fig. 3.4D,E). For instance, in the absence of input noise ($\sigma = 0$), the population's thresholds are all distinct from each other for all finite ranges of output noise except when $R \to 0$ (Fig. 3.4D). In the absence of output noise ($R \to \infty$), there is a critical value $\sigma_{\text{crit}} > 0$ at which the population transitions directly from all thresholds being distinct to all thresholds being equal (Fig. 3.4E). Note that for all these bifurcations the threshold differences change continuously, i.e. there are no jumps of optimal threshold values with varying noise.

Surprisingly, for the independent-coding channel I found a small range of input noise, $0.54 < \sigma < 0.6$, for which I observed a non-monotonic change in the number of distinct optimal thresholds when varying the output noise R (visible in the tail of red area indicating three distinct thresholds in Fig. 3.4A): for this range of σ, the optimal number of distinct thresholds with increasing output noise is 3-2-3-2-1, i.e. the optimal number of distinct thresholds changes from three thresholds being distinct, to two thresholds being distinct, to again three thresholds being distinct, to two thresholds being distinct before finally, the optimal solution is that all three thresholds are equal. I will treat this finding with more detail and within a greater context in Section 3.5.

Optimal thresholds of the lumped-coding channel

In comparison, for the lumped-coding channel, the bifurcations occur as the threshold differences at critical noise values change abruptly, i.e. discontinuously, when one noise source varies and the other remains fixed (Fig. 3.4G,H). Here, the system has an intermediate degree of redundancy, i.e. two thresholds being distinct, for a large range of noise values, and the transition from one to three distinct thresholds is not simultaneous as either noise vanishes. Rather, the discontinuous threshold jumping at each bifurcation becomes continuous (Fig. 3.4I), as normally seen for the independent-coding channel, or partly continuous (Fig. 3.4J). In Chapter 4 I will discuss in great detail the cause and implications of these differences between the two channel types. My results in the cases of one noise source vanishing for the lumped-coding channel agree with two previous studies, where a lumped-coding channel was studied with only output noise (Fig. 3.4I) [86], or with only input noise (Fig. 3.4J) [85]. In Section 3.4.3 I will show that my results in these limit cases are in fact equivalent to previous studies. My results are also consistent with previous studies for small populations of two neurons and only one source of noise [29,59,62], large populations with only output noise [37] and two-neuron populations with multiple noise sources [76].

3.4.2 Results for larger populations (independent-coding channel)

Remark: Two figures of this subsection (Fig. 3.5A, and 3.6B) have been previously published in the supplementary material of [1] where they were also briefly mentioned in the main text. Here, I show these two figures again and present and discuss them in a larger context.

The smaller computational load of calculating the mutual information of the independent-coding channel allowed me to study optimal thresholds for larger neural populations with up to six neurons. There, optimal thresholds, in general, show the same pattern of full redundancy (i.e. all thresholds are equal) at high noise levels and no redundancy (i.e. all thresholds are distinct) at low noise levels.

For four neurons, the transition from full to no redundancy with decreasing noise levels happens gradually through three transitions in the presence of both input and output noise (Fig. 3.5B,C). In the case of only output noise, the thresholds are fully distinct for all finite noise values (Fig. 3.5D), while for only input noise two transitions happen simultaneously at the same noise value (Fig. 3.5E).

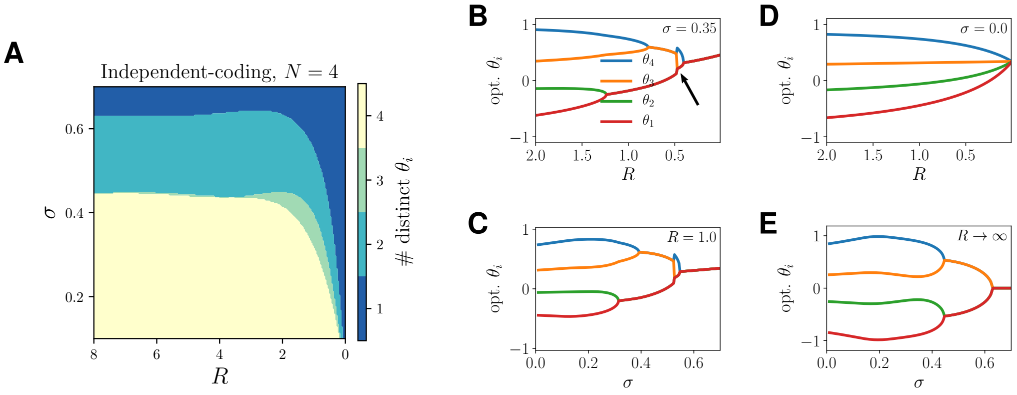


Figure 3.5: Optimal thresholds for a population of four neurons (independent-coding channel only). A. The optimal number of distinct thresholds depend on input noise σ and output noise R. Adapted from [1]. **B.** The optimal thresholds as a function of output noise for a fixed value of input noise ($\sigma = 0.35$). The arrow indicates a discontinuous threshold bifurcation. **C.** The optimal thresholds as a function of input noise for a fixed value of output noise ($R = 1$). **D.** The optimal thresholds as a function of output noise in the limit of no input noise ($\sigma = 0$). **E.** The optimal thresholds as a function of input noise in the limit of vanishing output noise ($R \to \infty$).

A phenomenon that is not seen for the independent-coding channel with only three neurons is that of discontinuous threshold bifurcations, i.e. where optimal thresholds suddenly "jump" from one value to a very different value with changing noise. These discontinuous threshold bifurcations appear near the bifurcation to full redundancy, interestingly while the optimal number of distinct thresholds remains two (Fig. 3.5B, arrow). One can argue that having three equal thresholds and a fourth being distinct, i.e. $\theta_4 \neq \theta_3 = \theta_2 = \theta_1$, means more redundancy (or less diversity) than having two times two equal thresholds, i.e. $\theta_4 = \theta_3 \neq \theta_2 = \theta_1$, even though the number of distinct thresholds is two in both cases. With that argument, this behavior of a discontinuous threshold jump of θ_3 fits very well into the general pattern of increased redundancy with increasing noise. Nevertheless, I will treat this phenomenon of discontinuous threshold jumps with more detail in Section 3.4.4

since they imply interesting consequences.

For six neurons, the basic pattern already described for three and four neurons stays the same for high output noise (Fig. 3.6A,C,D,I). However, more irregularities appear for lower output noise (Fig. 3.6B,E-H,J,K). The reason for these irregularities is that for some noise ranges the two middle thresholds, θ_3 and θ_4, get so close

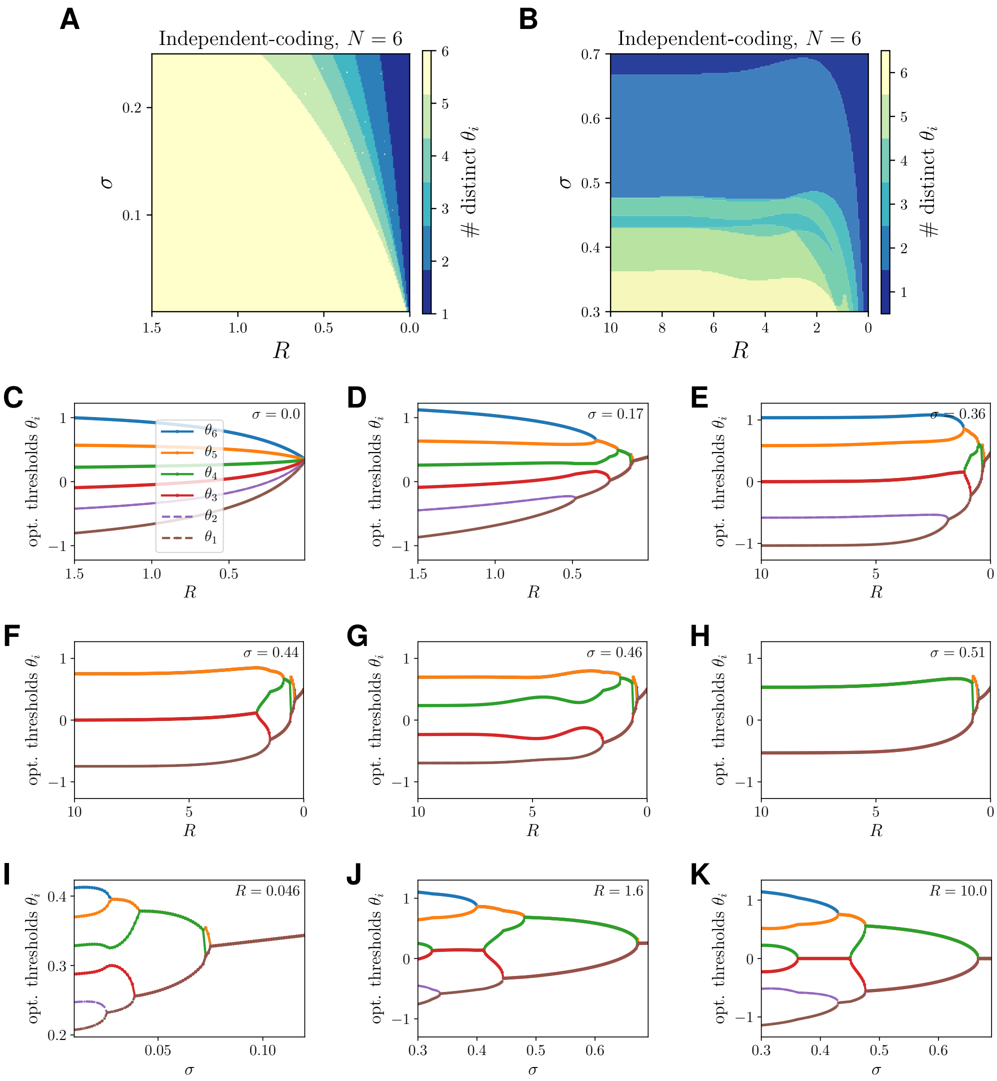


Figure 3.6: Optimal thresholds for a population of six neurons (independent-coding channel only). **A.** Optimal number of distinct thresholds depending on input noise σ and output noise R for small σ and a small range of R. Single beige dots in the darker regions are due to local maxima. **B.** As A but with a wide range of σ and large R. Adapted from [1]. **C-H.** Optimal thresholds for slices of fixed input noise. **I-K.** Optimal thresholds for slices of fixed output noise.

to each other that is more optimal that they have equal values instead of having distinct, but very similar values (Fig. 3.6E,F,J,K). This behavior is also a form of non-monotonic change in the number of distinct optimal thresholds, but again, these non-monotonicities will be treated in Section 3.5.

In conclusion, I can say that at least for the independent-channel the basic pattern of transition from full threshold diversity to full redundancy through subsequent bifurcations also remain valid for larger populations. However, with larger populations more and more special phenomena like discontinuous bifurcations and non-monotonic threshold behavior occur and the patterns of optimal thresholds become less clean.

3.4.3 Comparing results in the limit of one noise source with results from the literature (lumped-coding channel)

For the lumped-coding channel and one noise source only, the calculations are much less involved and have been described previously: McDonnell et al. looked at input noise only [85] and Nikitin et al. looked at output noise only [86]. Using their methods,[8] I reproduced their results and compared the results of my calculations with one noise source being very small to their respective results.

First I compared my result with very small input noise with Nikitin et al. For $N = 3$ and very small input noise of $\sigma = 0.01$ my calculations lead to very similar optimal thresholds (Fig. 3.7A) and maximized mutual information (Fig. 3.7B) as Nikitin et al.'s calculations. The exceptions are one local maximum shortly before the first threshold merging and small deviations in maximized information for large R. The former is likely due to a local maximum which would be resolved with a grid of initial optimization values, while the latter is due to the fact that $\sigma = 0.01$ is considerably larger than $\sigma = 0$. However, it was not possible to run a grid of initial values or to further decrease σ, since both a large R and a small σ are very expensive with $N = 3$ neurons.[9]

Then I compared my result with very small output noise with McDonnell et al. Since very small output noise here means $R \gtrsim 50$ (with $R = 35$, for example, there are still considerable differences, not shown), I could only do this comparison for

[8] For both of their methods, the underlying approach of numerically optimizing the mutual information is similar to that described in my work. The method of McDonnell et al. is analog to that in Sec. 3.1.2 but with less summation terms due to missing output noise. The method of Nikitin et al. allows for fast numerical calculation of the mutual information since no numerical integration has to be performed, due to the error functions (see Eqs. 3.12-3.14) becoming binary. However, I slightly had to modify their setup: in their work, they also optimize the optimal weighting of the single channels before lumping them. I set these weights to $1/N$ since I have an equal weighting of all channels.

[9] The result shown in Figure 3.7A,B took more than three days of computing on a cluster with 32 cores.

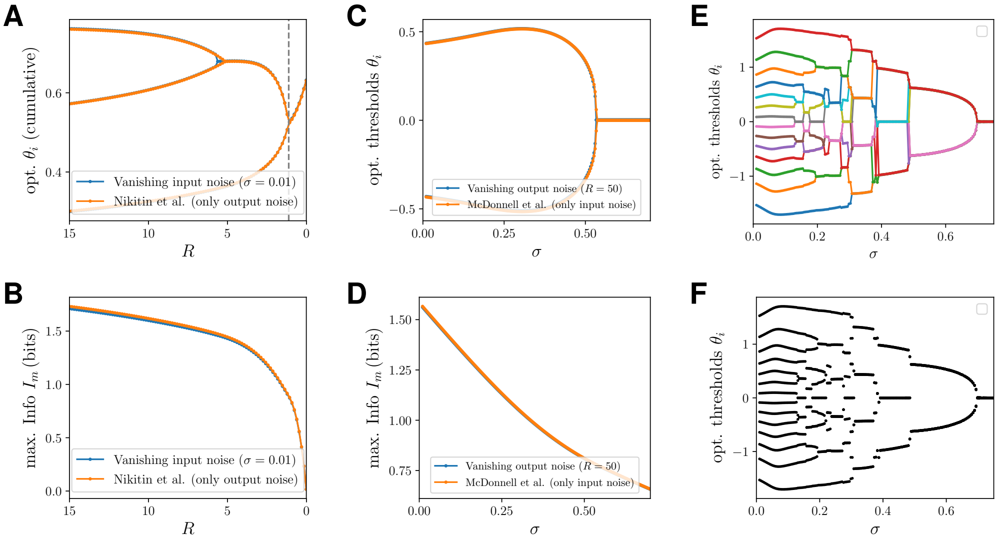


Figure 3.7: The results using the lumped-coding channel in the limit of one noise correspond to results from the literature. A, B. The lumped setup ($N = 3$ cells) with very small input noise ($\sigma = 0.01$) is compared to Nikitin et al., who looked into a lumped setup with only output noise [86]. **A, B.** The lumped setup ($N = 3$ cells) with very small input noise ($\sigma = 0.01$) is compared to Nikitin et al., who looked into a lumped setup with only output noise and can calculate the mutual information without numerical integration [86]. **A.** Optimal thresholds depending on output noise. The dashed vertical line denotes the first threshold split calculated analytically for the lumped-coding channel without input noise [86]. **B.** Maximized mutual information depending on output noise. **C, D.** The lumped setup (only $N = 2$ cells since numerical calculations become not feasible for $R \gg 10$ and $N = 3$) with very small output noise ($R = 50$) is compared to McDonnell et al., who looked into a lumped setup with only input noise [85]. Note that the curves lay on top of each other. **E.** Optimal thresholds in the lumped setup with only input noise and large neuron number ($N = 15$), reproducing the result of McDonnell et al. but plotting it with continuous lines of different colors. That way, local maxima and threshold switches can be easily spotted. **F.** The same result as in E but in style McDonnell et al. plotted optimal thresholds (see their Fig. 3 in [85]). Due to unconnected dots one often does not see the existence of local maxima, threshold switching, or how exactly which thresholds bifurcate.

$N = 2$. For this case, both the optimal thresholds and the maximum information are in accordance with McDonnell et al. (Fig. 3.7C,D). Therefore I conclude that my way of calculating the mutual information in the case of a lumped setup with both additive input noise and Poisson output noise and my way of optimizing the optimal thresholds is consistent with previously published research.

As a side note I want to point out that McDonnell et al.'s way of plotting their optimal thresholds with black dots only (Fig. 3.7F) is potentially concealing the existence of suboptimal thresholds due to local maxima: when comparing their way of plotting to my way of plotting thresholds in color and with connecting lines (Fig. 3.7E) one sees that many irregularities can be concealed.

3.4.4 Avoiding suboptimal thresholds in the case of local maxima

With optimization problems, there is always the general possibility that there exist one or more local optima in addition to a global optimum. Classical optimization algorithms in general just find the (local) optimum in whose proximity they started the search. Now that the reader has become familiar with how optimal number of distinct thresholds and the optimal threshold values behave with changing noise values, I am going to use some examples to explain how I avoided local maxima.

For $N = 2$ neurons it is still possible to plot the information landscape $I_m(\theta_1, \theta_2)$ as a heatmap and with contours to get an impression if there are local maxima. Already for $N = 3$ this is not trivial anymore. A simple approach is to use a grid of initial threshold values and for each element of the grid perform a local optimization. This approach also gets computationally involved for larger N as the number of local optimizations that have to be performed scales with the power of N. Therefore, for high-dimensional optimizing one usually uses global optimization algorithms, which are designed to find the global maximum in a landscape of many local maxima [96].

There exist a large zoo of global optimization algorithms, which are often specifically designed for the properties of the specific landscape. These properties include the number and distributions of local maxima, the size and range of the basins of the maxima, how much the peaks of the maxima actually differ, etc. Fortunately, it turned out that my information landscapes are usually very concave[10], especially for smaller populations: for the independent-coding channel it was impossible to find local maxima for $N = 2$ and still not easy[11] for $N = 3$. For $N = 4$ and $N = 6$, local maxima regularly appeared, but were easy to spot (described below). For the lumped-coding channel, local maxima only appeared at threshold bifurcations, and even there the whole information landscape had exactly one local maximum. To be able to comprehensively describe the details, I give an overview of four different types of local maxima that I encountered and describe them step by step (Tab. 3.1).

[10] I.e. they are positively curved – like a parabola – almost everywhere. Mathematically, this is the case when the second derivative of the information with respect to a threshold is negative for all thresholds at every point [100].

[11] I tried a grid of initial vectors to find local maxima, but the only initial threshold vector that gave more than a handful of local maxima (in most cases it was just zero) out of 40 000 noise combinations was the null-vector, i.e. $\vec{\theta} = (0, 0, 0)$. Even in this case, local maxima are the absolute minority (see Fig. 3.10).

Table 3.1: Overview of the different types of local maxima. "θ bifu." means that the local maximum happens near (in terms of noise values being close to critical noise values) bifurcation of optimal thresholds or are directly related to it. "$\#d_\theta$ eff." means that the optimal threshold diversity, $\#d_\theta$, is effected and thus differs between the local and the global maximum; these are very easy to spot in the color maps showing $\#d_\theta(R, \sigma)$ (see Fig. 3.10A) and thus called "intrusive". The columns "Indep." and "Lump." show that some types of local maxima do only appear for a minimum population size or do not appear at all for the respective channel types. The last column specifies how the respective type of local maximum can be avoided: either by a grid of initial threshold values θ_0 or by adapting θ_0 to the optimal thresholds from slightly different noise values (see Fig. 3.9A-C)

Type	**Example**	**θ bifu.**	**$\#d_\theta$ eff.**	**Indep.**	**Lump.**	**Avoiding**
(1) Discont. bifurcation	Fig. 3.4G,H, 3.9B,C	yes	yes	not seen	always	adapt θ_0
(2) Discont. switching	Fig. 3.5B,C, 3.9D	yes	no	$N \geq 4$	not seen	adapt θ_0
(3) "Unintrusive" random	Fig. 3.9E	no	no	rarely	rarely	adapt θ_0
(4) "Intrusive" random	Fig. 3.10A-C	no	yes	$N \geq 3$	not seen	grid of θ_0

Overview of different types of local maxima

Remark: Some panels of Figure. 3.8 in this subsection have appeared in a similar form in a different context in [1]. Here, I use adapted and extended versions of these panels to discuss results in a different context.

First, there are local maxima due to discontinuous threshold bifurcations which I call type (1) and (2). They necessarily appear since every discontinuous threshold bifurcation implies a switch from a local to a global maximum (Fig. 3.8A,B; also treated in more detail in Fig. 4.4 in Sec. 4.1). In contrast, continuous threshold splits do not involve a switch from a local to a global maximum (or vice versa) and thus not necessarilly imply the presence of a local maximum (Fig. 3.8C,D).

At type (1) local maxima, the optimal number of distinct thresholds, $\#d_\theta$, actually changes (see Fig. 3.4G,H). To reliably avoid confusing type (1) local maxima as global maxima, optimal thresholds are obtained for each noise value by using two different initial threshold vectors $\vec{\theta}$ in the optimization process. First, by using the optimal thresholds from the slightly larger noise value as initial values, $\vec{\theta_0} \downarrow$, and second, by using the optimal thresholds from the slightly smaller noise values, $\vec{\theta_0} \uparrow$ (Fig. 3.9A). Then for each noise level two optimization procedures are performed, one with $\vec{\theta_0} \downarrow$ as initial threshold vectors and the other with $\vec{\theta_0} \uparrow$. The optimization procedure with $\vec{\theta_0} \uparrow$ as initial conditions reliably finds the maximum that is global for low noise values, which then becomes local at the critical noise value, and which then vanishes eventually for further increased noise (see again Fig. 3.8A,B). The optimization procedure with $\vec{\theta_0} \downarrow$ as initial conditions reliably finds the maximum that is global for high noise values, which then (with decreasing noise) becomes local at the critical noise value, and which then vanishes eventually for further decreased noise. By comparing the two maxima for each noise value, one can identify the

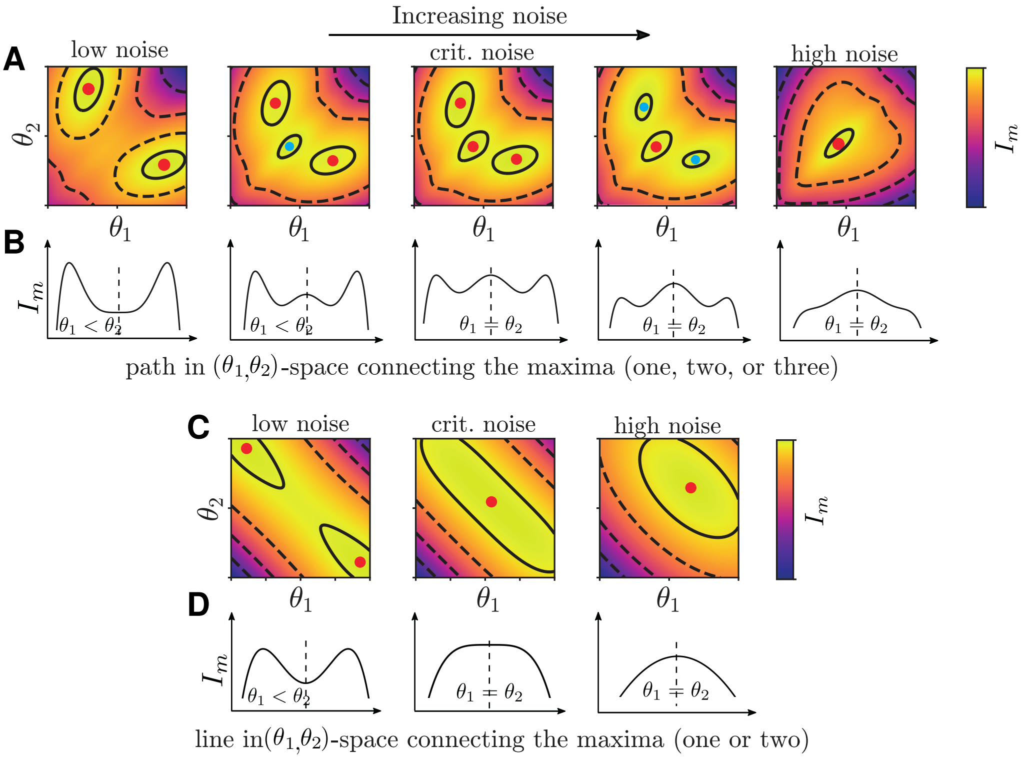


Figure 3.8: Discontinuous threshold bifurcations imply at least one local maximum. A. Information landscapes for continuous threshold bifurcations. Each panel qualitatively shows the information landscape $I_m(\theta_1, \theta_2)$ for a given noise value, from left to right the noise value increases. Global maxima are shown in red, local maxima in blue. At noise levels close to the critical noise level, there exist at least one local maximum. At the critical noise level, the local maximum becomes global and vice versa (middle panel). **B.** Schematic slices through the information landscape to visualize the switch from local to global maxima. **C.** As in A but for continuous threshold bifurcations. In general, there exist no local maxima. **D.** Schematic slices through the information landscape.

global and the local maximum with the respective optimal and suboptimal thresholds at each noise value (Fig. 3.9B). The information differences between local and global maxima are shown in Figure 3.9C.

At type (2) local maxima, there are also discontinuous bifurcations of optimal thresholds, however, the optimal number of distinct thresholds, $\#d_\theta$, does not change (see Fig. 3.5B, arrow; and Fig. 3.9D), since one or more optimal thresholds just “switch sides”: in Fig. 3.5B, for example, with decreasing R, θ_3 switches from being equal to θ_4 to being equal with θ_1 and θ_2; during that switch the number of distinct optimal thresholds does not change and remains constant at $\#d_\theta = 2$. Type (2) local maxima cannot be identified using the color maps showing $\#d_\theta(R, \sigma)$ (like Fig. 3.5A). Instead, they can only be seen when plotting optimal thetas in de-

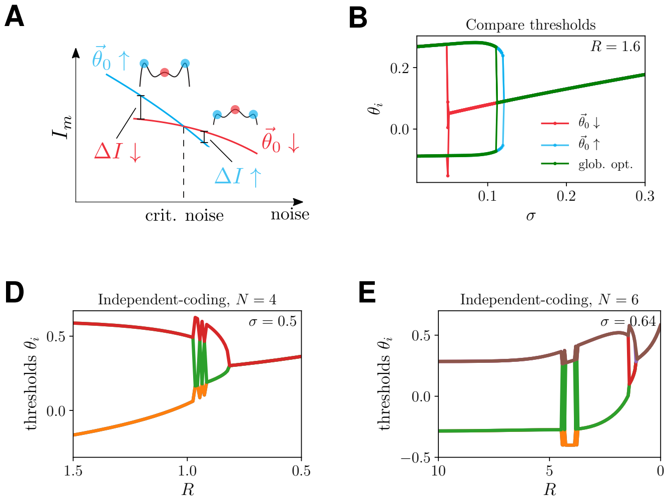


Figure 3.9: Different examples of local maxima due to discontinuous threshold bifurcations. A. Schematic showing how the mutual information of both local and global maxima depends on noise. At discontinuous threshold bifurcations, the local maximum becomes global and vice versa. Two different initial threshold vectors $\vec{\theta}_0$ are used for the optimization at each noise value: $\vec{\theta}_0 \downarrow$ (red) is the optimal threshold vector from the slightly larger noise value and $\vec{\theta}_0 \uparrow$ (blue) is the optimal threshold vector from the slightly smaller noise. The information difference between local and global maximum is denoted as $\Delta I \downarrow$ and$\Delta I \uparrow$, respectively. **B.** Thresholds resembling both the local and global maxima around a discontinuous threshold bifurcation (lumped channel with $N = 3$ neurons). Blue and red lines show the optimized thresholds using $\vec{\theta}_0 \downarrow$ and $\vec{\theta}_0 \downarrow$ as initial values for optimization. The green line shows the globally optimal thresholds for each σ, which is obtained through choosing for each σ the thresholds which lead to higher information (see next panel). **C.** Information difference between optimal thresholds and the ones obtained in B by using $\vec{\theta}_0 \downarrow$ and $\vec{\theta}_0 \uparrow$ as initial threshold values. **D.** An example of local maxima around discontinuous threshold bifurcations using the independent-coding channel ($N = 4$ neurons). For a range of output noise values R the optimizer jumps between the local and the global maximum. **E.** An example of local maxima not related to threshold bifurcations. Independent-coding channel with $N = 6$ neurons.

pendency of one noise source while fixing the other one (e.g. Fig. 3.5B). Since they only appear for $N \geq 4$, I did not spend too much time on techniques for reliably spotting or avoiding them. However, in principle they can be avoided by the same technique of using adapted initial threshold vectors as used for type (1) local maxima, i.e. using optimal threshold vectors from both lower and higher noise values as initial vectors.

Type (3) local maxima are not related to any bifurcations of optimal thresholds and the optimal number of distinct thresholds does not change (Fig. 3.9E). Thus they are difficult to spot using color maps showing $\#d_\theta(R, \sigma)$. In addition, they seem to appear randomly, at least I have not found a clear pattern. Thus, for type (3) local maxima one again has to look at the slices where the optimal thresholds are plotted against one noise level while the other one is fixed. Luckily they do appear very rarely and can be reliably avoided using adapted initial threshold vectors for optimization.

Type (4) local maxima are also not related to any bifurcations of optimal thresholds but for them, the optimal number of distinct thresholds is effected, so they can easily be spotted using color maps showing $\#d_\theta(R,\sigma)$ (Fig. 3.10A). They also seem to appear randomly. I have not seen them for the lumped channel, and for the independent channel, they only appear when using the null-vector as the initial threshold vector for optimization or for $N=6$. They can in most cases be avoided by using initial threshold vectors as for the previous three types, however, in some cases for $N=6$ they can only be reliably avoided by using a grid of initial threshold vectors for optimization.

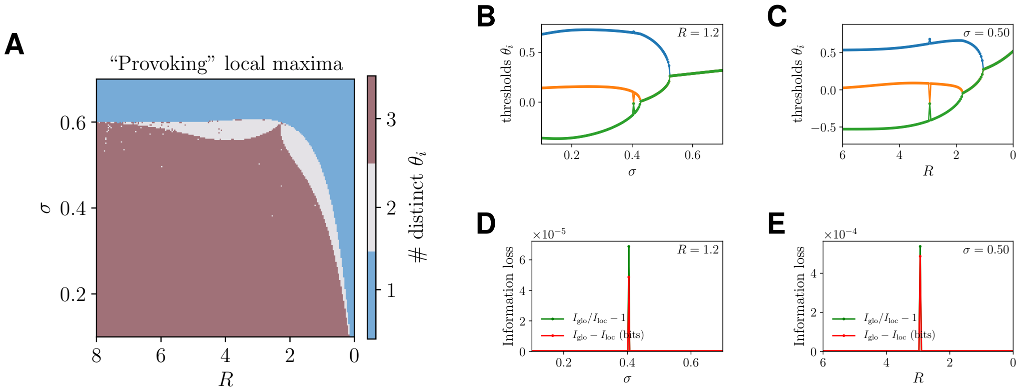


Figure 3.10: Obtaining a few local maxima by changing the initial values fed to the optimization algorithm. To find local maxima for the independent-coding channel and $N=3$, initial threshold values being used for the optimization process are changed to $\vec{\theta_0}=\{0,0,0\}$ (instead of $\vec{\theta_0}=\{-0.4,0.2,0.8\}$). **A.** Optimal number of distinct thresholds (compare to Fig. 3.4A). The single white dots in the red space and the single red dots in the white space represent noise combinations for which the optimization algorithm ran into a local maximum. **B, C.** Two slices through A for which suboptimal thresholds occur. These are clearly visible as single threshold jumps. **D, E.** Corresponding information loss at the local maximum versus the global one. The information difference and the information ratio are very close to zero or one, respectively.

Conclusion

For the different types of local maxima there exist techniques to reliably spot these suboptimal thresholds. Mostly, however, the local maxima only appear for $N\geq 4$ and thus do not play a big role for most of my work. Besides, the information difference between local and global maxima are often rather small, for example in the order of 10^{-4} bits for $N = 3$ (Fig. 3.10D,E) and even smaller for $N > 3$ (usually smaller than 10^{-6} for $N = 4$, data not shown). Thus, the question arises if biological systems really go for the global maximum or if they might not also settle for a local maximum that is almost identical to the global maximum in terms of efficiency. This question will be treated in Section 4.2.

3.4.5 Comparing two local optimization algorithms

In Section 3.2.2 I promised to show that for the lumped-coding channel the BFGS which is based on estimating the gradient of the information landscape provides the same results as the NM optimization algorithm which is based on the simplex method. For my optimizations, the BFGS algorithm is about two times faster than the NM algorithm, however, the BFGS algorithm relies on inverting the Hessian matrix what leads to large numerical errors in the case of continuous threshold bifurcations (see Sec. 4.3). Since the threshold bifurcations are discontinuous for the lumped-coding channel it should be fine to use it there. Nevertheless, I show that for the lumped-coding channel the two optimizers provide indeed the same – or at least extremely similar – results of maximized mutual information and optimal thresholds: the difference in maximized information between the two optimizers is on the order of 10^{-8} bits for three neurons and a grid of 200 times 200 noise combinations (Fig. 3.11A), and also the information ratio deviates less then 10^{-7} from one (Fig. 3.11B). Besides, the sum of absolute threshold differences, $\sum_i^{N=3} |\Delta\theta_i|$, is on the order of 10^{-3} (Fig. 3.11C). These results demonstrate, that I can indeed use the speed advantage of the BFGS algorithm for the lumped-coding channel.

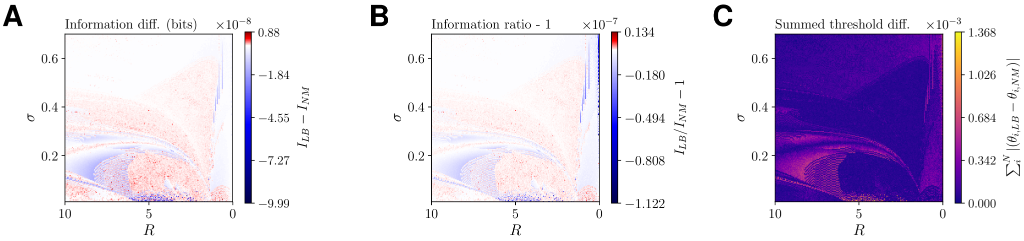


Figure 3.11: Comparing information and thresholds when using the L-BFGS algorithm ("LB") compared to the Nelder-Mead algorithm ("NM") for optimizing the lumped-coding channel ($N = 3$). A. Color coded is the absolute information difference in bits for each noise combination when maximizing mutual information using the two different optimizers (note the 10^{-8} scale). **B.** Similarly to A, here is shown the information ratio minus one in order to demonstrate that the information ratio is very close to one for all noise level combinations. **C.** The sum of all three (since $N = 3$) absolute threshold differences for each noise level combination is color-coded (note the 10^{-3} scale).

3.5 Non-monotonous behavior of optimal thresholds

Now I will treat the unexpected results of non-monotonic behavior of optimal thresholds in the case of the independent-coding channel. There, instead of the expected behavior of optimal number of distinct thresholds decreasing with increasing noise, the optimal number of distinct thresholds is a non-monotonous function of noise. In this section, I will first describe in detail three different classes of non-monotonicity in the optimal number of distinct thresholds. Since some of this non-monotonous behavior is very counter-intuitive and lacks a comprehensive explanation I have put considerable effort into excluding potential errors or numerical instabilities. Thus, I will describe the extensive approaches to exclude numerical imprecisions, local maxima and non-analytical behavior in the functions of which the mutual information is composed of. Finally, I will show that a non-monotonic behavior of optimal threshold *differences* have occurred already in some previous studies and that this non-monotonic behavior of threshold differences is closely related to the non-monotonic behavior of the optimal number of distinct thresholds. For brevity, I will from now on use the term “threshold diversity” for “number of distinct thresholds”.

3.5.1 Three classes of non-monotonicity in the optimal number of distinct thresholds

Remark: Some figures of this subsection (Figs. 3.12B,G) have been published in a similar form in the supplementary material of [1] and have been briefly mentioned in the main text of [1]. Here, I show these two figures again and present and discuss them in a larger context.

All of the non-monotonicities that I found can be classified into one of the three groups which I call “threshold ribbons”, “threshold switching”, and “threshold splitting”.

Threshold ribbons

Threshold ribbons are arguably the most pronounced and unexpected form of non-monotonic threshold behavior. In the most simple form, two thresholds which are equal at a certain noise level will first split and then merge again with increasing noise level. In Figure 3.12A, for example, at a certain value of input noise the optimal threshold diversity with increasing output noise is “2-1-2-1”, which means that first having two distinct thresholds is optimal, then having two equal thresholds (i.e. one distinct threshold) is optimal, before having two distinct optimal thresholds and then finally just one again. Such a ribbon behavior also exists for populations with more than two neurons: for example, in a population of three neurons, the optimal threshold diversity is 1-2-1 with increasing output noise at a certain input noise level (Fig. 3.12C). For a population of four neurons, there

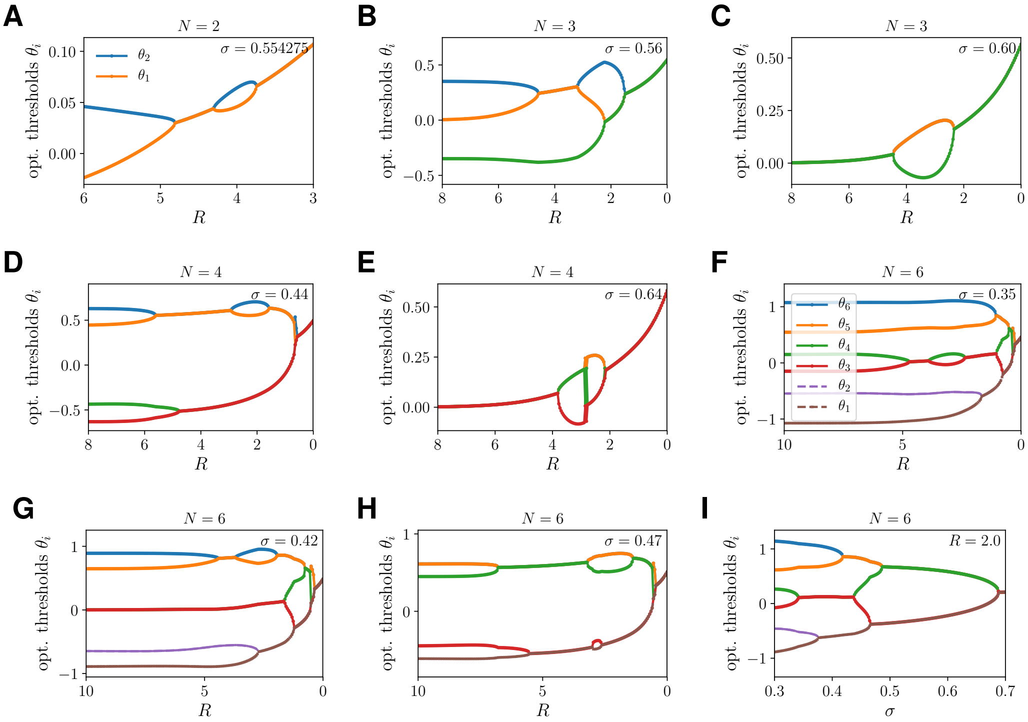


Figure 3.12: For certain values of one noise source, optimal number of distinct thresholds change in a non-monotonic way with the other noise source (independent-coding channel only). **A.** “Threshold ribbons”: For $N = 2$ neurons and varying output noise R (larger R means less output noise) the optimal thresholds show a sequence of three bifurcations, leading to a peculiar ribbon of optimal thresholds for fixed input noise σ). **B.** “Threshold switching”: With increasing output noise the optimal number of distinct thresholds changes from three to two, to again three, then again to two, and finally to one ($N = 3$). **C.** A threshold ribbon in the case of $N = 3$ neurons. **D.** A ribbon of two thresholds splitting and merging again, without any effect on the other two thresholds ($N = 4$). **E.** A ribbon in which all four thresholds participate. There also happens a discontinuous threshold switch inside the ribbon which is not affecting the optimal number of distinct thresholds. **F-H.** Ribbons in the case of $N = 6$. **I.** Threshold switching in the case of $N = 6$ along the input noise axis when fixing the output noise value ($R = 2$). The optimal number of distinct thresholds is 6-5-4-3-4-3-2-1. The panels B, C, and F are from [1].

exist two different ribbon behaviors, namely one with optimal threshold diversity of 4-3-2-3-2-1 at $\sigma \approx 0.44$ (Fig. 3.12D), and one with 1-2-1 at $\sigma \approx 0.64$ (Fig. 3.12E). Note that in the latter example the non-continuous switching of optimal thresholds comes from a local maximum becoming the global maximum and vice versa with changing R (a type (2) local maxima, see Tab. 3.1 in Sec. 3.4.4). For a population of six neurons, numerous ribbons appear at various σ, but here I just list three examples: one with optimal threshold diversity of 6-5-6-5-4-... (Fig. 3.12F), one with 5-4-5-4-3-... (Fig. 3.12G), and one with 4-3-2-3-4-3-2-1 (Fig. 3.12H).

Ribbons in general only appear for a fixed input noise and with varying output noise, i.e. they, in general, do not appear for fixed output noise and varying input noise.[12] Usually, the range of input noise for which ribbons appear is very small but becomes larger for larger N. The ribbon for $N = 2$, for example, appears only in the range of $\sigma \approx [0.55426, 0.55431]$, while the ribbon of $N = 4$ appears in the range of $\sigma \approx [0.434, 0.447]$. In general, it can be said that the larger N, the more ribbons appear, the more pronounced they are, and the larger the range of input noise σ is at which they appear.

Threshold switching

The non-monotonic behavior I call threshold switching is not as peculiar as the ribbons. With Gaussian input noise it only occurs for a population of three neurons, however there it occurs in a wide range of input noise, namely $\sigma \approx [0.54, 0.60]$ (Fig. 3.12B).[13] During threshold switching the following happens: first, at lower output noise, the upper two thresholds being equal is optimal, and then, at somewhat higher output noise, the lower two thresholds being equal is optimal. Between these two states, there is a continuous transition from one state to the other, at which all three thresholds being distinct is optimal. Thus, the optimal number of distinct thresholds with increasing output noise is 3-2-3-2-1. The reason that threshold switching does not appear for $N > 3$ could be that for $N > 3$ the threshold switching happens discontinuously, for which there is no intermediate regime with increased threshold diversity necessary (Fig. 3.12E and Fig. 3.5B, arrow). Still, it remains unclear for $N = 3$ why, (1) it is optimal that the upper two thresholds are equal for relatively low output noise, while it is optimal that the lower two thresholds are equal for somewhat larger noise, and why, (2) the switching of the middle threshold from being equal to the upper threshold to being equal to the lower threshold happens continuously and not in a discontinuous manner as it is the case for $N = 4$.

Threshold splitting

This class contains the form of non-monotonic threshold behavior which is not covered by the first two classes. Threshold splitting is very prevalent for populations with more than four neurons. For six neurons, it happens for large ranges of input noise and even – in contrast to the previous two classes – when fixing the output noise and varying the input noise. This form is characterized by the fact that the two middle thresholds, θ_3 and θ_4, get so close to each other that it is more optimal that they have equal values instead of having distinct, but very similar values (Fig. 3.6E,F,J,K and Fig. 3.12F,G,I). The reason why they get so close to

[12] However, they can be seen for fixed output noise when using not a Gaussian input noise distribution but an input noise distribution with lower kurtosis (Fig. 3.19 in Sec. 3.6.1).

[13] For the case of non-Gaussian noise, it also happens for $N = 4$ (Fig. 3.19D,E in Sec. 3.6.1).

each other in the first place remains unknown. The optimal threshold diversity can behave like 5-4-3-4-3-2-1 (Fig. 3.6E, though it is difficult to verify in this plot size), like 3-4-3-2-1 (Fig. 3.6F), or like 6-5-4-3-4-3-2-1 (Fig. 3.12I).

3.5.2 Excluding potential errors that cause non-monotonic results

Since the non-monotonicities described above are very unintuitive and I have not found a comprehensible explanation, I have invested a lot of efforts into verifying the results above. In principle, three possible sources of error exist that could cause systematically wrong results: (1) the optimization procedure is faulty or ends up in local maxima, (2) the numerical integration introduces significant errors, and (3) during the numerical calculations of the mutual information numerical instabilities are introduced, e.g. due having functions that cause non-analytic terms, like poles. Now, I will show that neither of them seems to be the case.

3.5.2.1 Verifying the optimization procedure

Eight different optimization algorithms of the Scipy package [101] were used to find optimal thresholds for three and four neurons at an input noise value at which the threshold diversity showed non-monotonic behavior (Fig. 3.13A,B). Remarkably, all algorithms reliably reproduce the non-monotonic behavior of threshold switching (Fig. 3.12A) and the threshold ribbon (Fig. 3.12B). The basin-hopping algorithm is a global optimizer, while all other algorithms are local optimizers, which rely on gradient descent (BFGS, L-BFGS, SLSQP, conjugated gradient) or use other techniques (Nelder-Mead, Powell's method, COBYLA) [96]. Except for the basin-hopping, the Nelder-Mead, and the (L-)BFGS algorithm I did not look into how they function or choose specific adjustments. Additionally for $N = 4$, I performed a simple grid search for each $R \in \{7, 4, 2., 0.75, 0.1\}$ and $\sigma = 0.44$ with 150^4 grid elements in the threshold space of $[-0.8, 0.8]^4$ for each R; as well as Nelder-Mead optimization with a grid of initial threshold vectors with 33^4 grid elements in the threshold space of $[-0.8, 0.8]^4$ (data not shown). Both approaches are in accordance with the ribbon result in Fig. 3.13B. Additionally, I also implemented the case with $N = 3$ in Mathematica (Wolfram Research) where using the "Random Search" algorithm for $R \in \{6, 4, 2.75, 1.95, 1\}$ and $\sigma = 0.56$. It also gave results in accordance with Fig. 3.13A (data not shown).

Furthermore, for $N = \{2, 3, 4\}$ I performed optimizations where I constrained thresholds to be equal, i.e. I made it impossible that threshold ribbons occur.[14] Then I compared the information with the unconstrained thresholds (showing ribbons) to the constrained thresholds (showing no ribbons): For each N, the constrained thresholds showing no ribbons cause an information loss at the noise range in which the ribbons occur (Fig. 3.14).

[14] These calculations I performed with Matlab (MathWorks) as their *fmincon* function handles constrained optimization problems very well.

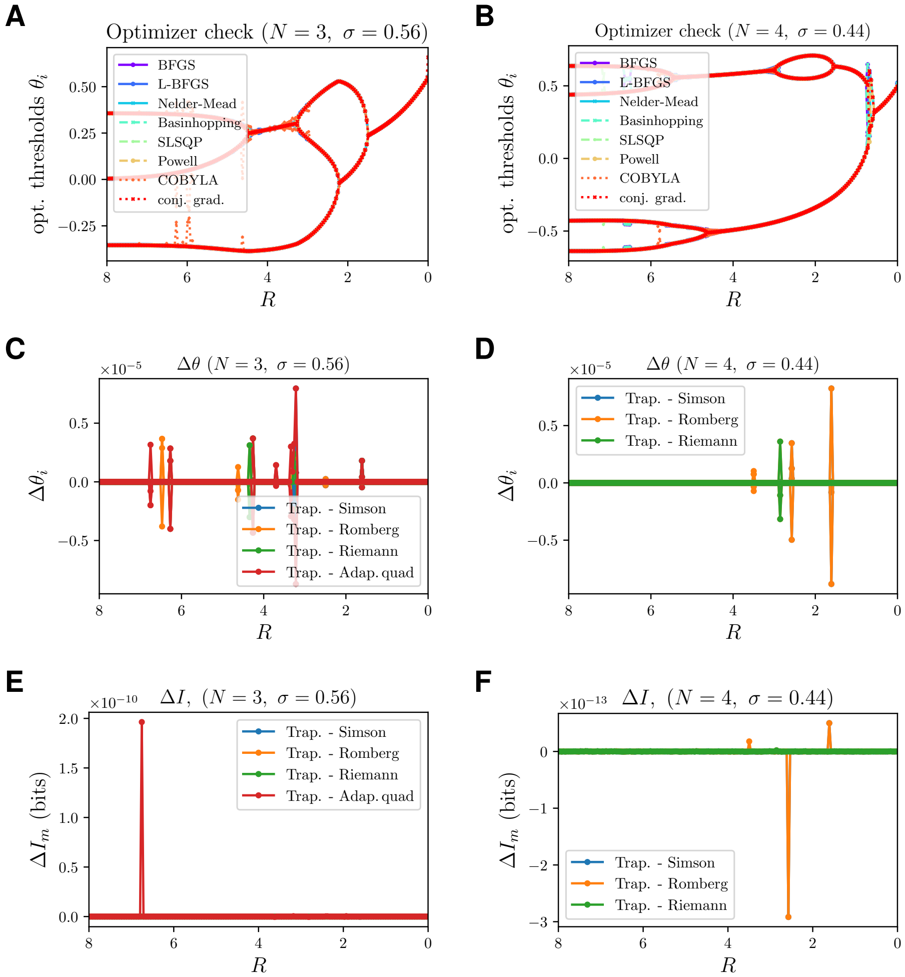


Figure 3.13: Checking if the non-monotonic threshold behavior is due to wrong optimization or numerical integration (left column: three neurons, right column: four neurons). **A, B.** Using different optimization algorithms produce the same qualitative result of non-monotonic threshold diversity. **C, D.** Using different integration algorithms lead to differences in optimal threshold in the order of 10^{-6}. **E, F.** The resulting information differences for the different thresholds obtained in (C, D) are in the order of 10^{-10} bits.

The facts that all optimizers lead to similar results, that performing a very dense grid search showed no hidden maxima, and that when optimizing while constraining thresholds to be equal leads to information loss, makes me confident that the unexpected results of non-monotonic threshold behavior are not caused by using optimization processes incorrectly.

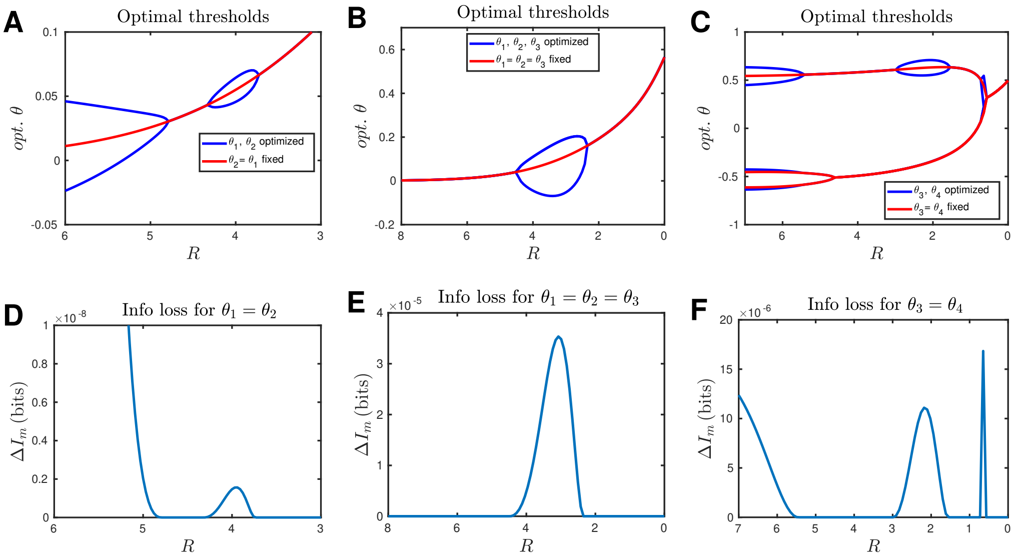


Figure 3.14: Doing a constrained optimization by forcing the thresholds participating in non-monotonic bifurcations to be equal results in information loss. **A.** "Optimal" thresholds in the case of constraining thresholds to be equal (red) compared to the unconstrained optimization (blue). Independent channel with $N = 2$ cells and input noise $\sigma = 0.554274$. **B.** As A but with $N = 3$ and $\sigma = 0.6$. **C.** With $N = 4$ and $\sigma = 0.44$, where just the upper two thresholds are constrained to be equal. **D-F.** The information loss in case of the constrained optimization compared to the unconstrained optimization from A-C.

3.5.2.2 Excluding numerical integration errors

Another potential error could have entered through the numerical integrations which have to be carried out in the calculation of the mutual information (Eqs. 3.5, 3.20). First, one has to integrate $P(\vec{k}|s)$ over s for all combinations of $k_i = \{0, 1\}$ and then one has to integrate $P(\vec{k}|s)\log_2(P(\vec{k}|s)/P(\vec{k}))$.[15] I used five different numerical integration methods: four from the Numpy package [102] which have a fixed step size (Riemann, Trapezoid, Romberg, Simson) and one adaptive integration algorithm from the Scipy package [101] where the step size is adapted. The adaptive integration was very slow due to large numbers of function evaluations and was thus only used for $N = 3$.

The assumption is the following: if the non-monotonic behavior is really introduced by numerical integration imprecisions, then different integration algorithms cause different imprecisions and thus lead to different thresholds. I compared optimized thresholds obtained by using the Trapezoid algorithm (which I used as a standard

[15] Since summing and integration can be swapped, the second integration can be done once for all combinations of k_i (i.e. 2^N times) before summing the integrals, or just once in total after having summed the terms into one integrand. To reduce the numerical integration error, I chose the latter possibility.

for all my calculations) with optimized thresholds obtained by using other integration algorithms at an input noise level at which non-monotonic threshold diversity occurs. I found that for several output noise values there is a threshold difference on the order of 10^{-5} but for most noise levels it is much smaller (Fig. 3.13C,D). In any case, the differences are so small, that the non-monotonic behavior still occurs. Unsurprisingly, then also the mutual information does not differ significantly, namely in the order of 10^{-10} bits and smaller (Fig. 3.13E,F). These results make me confident that imprecisions due to numerical integration errors are also not the reason for obtaining the unexpected threshold behavior.

3.5.2.3 Excluding numerical imprecisions due to non-analytic terms

Even though the equations which describe the calculation of the mutual information of the independent-coding channel do not indicate that any single term by itself is non-analytic (Sec. 3.1.1), in principle, there could also be the case of a pole, i.e. one variable becoming infinite. However, this pole-behavior could be hidden through something like

$$\lim_{x,y\to\infty} \frac{y}{x} \quad \text{or} \quad \lim_{x,y\to 0} \frac{y}{x}, \tag{3.31}$$

which still would cause numerical instabilities. To check for these hidden poles, I took the case of $N = 2$ neurons with the threshold ribbon at $\sigma = 0.554274$ (Fig. 3.12A). For each combination of output variables $\{k_i\}$ there exist a separate term of the mutual information (Eq. 3.20) which for $N = 2$ is given as

$$I_{k_1,k_2} = \int_s P_s(s)\, P_{k_1,k_2}(s) \log_2 \left(\frac{P_{k_1,k_2}(s)}{B_{k_1,k_2}} \right) \mathrm{d}s \tag{3.32}$$

where

$$P_{k_1,k_2}(s) = \prod_i^{N=2} P_{k_i}(s), \tag{3.33}$$

and

$$B_{k_1,k_2}(s) = \int_s \mathrm{d}s P_s(s) P_{k_1,k_2}(s). \tag{3.34}$$

For ease of notation, I just write I_{kk}, $P_{kk}(s)$, and B_{kk}. Note that each $k_i \in \{0, 1\}$ and thus there are 2^2 different I_{kk}, P_{kk}, and B_{kk}. Possible numerical instabilities arise when one or more B_{kk} or $P_{kk}(s)$ go to zero. In Fig. 3.15(D-F) I show how B_{kk} depends on output noise for specific values of input noise. I chose values of input noise for which the non-monotonic threshold behavior occurs ($\sigma = 0.554274$, Fig. 3.15B) as well as one value slightly below and above ($\sigma = 0.5542$ and $\sigma = 0.55435$, Fig. 3.15A and C, respectively). It is not well visible in the plots, however, I confirmed that all B_{kk} are at least 0.01 for all $R > 1$ and any σ. Similarly, all I_{kk}

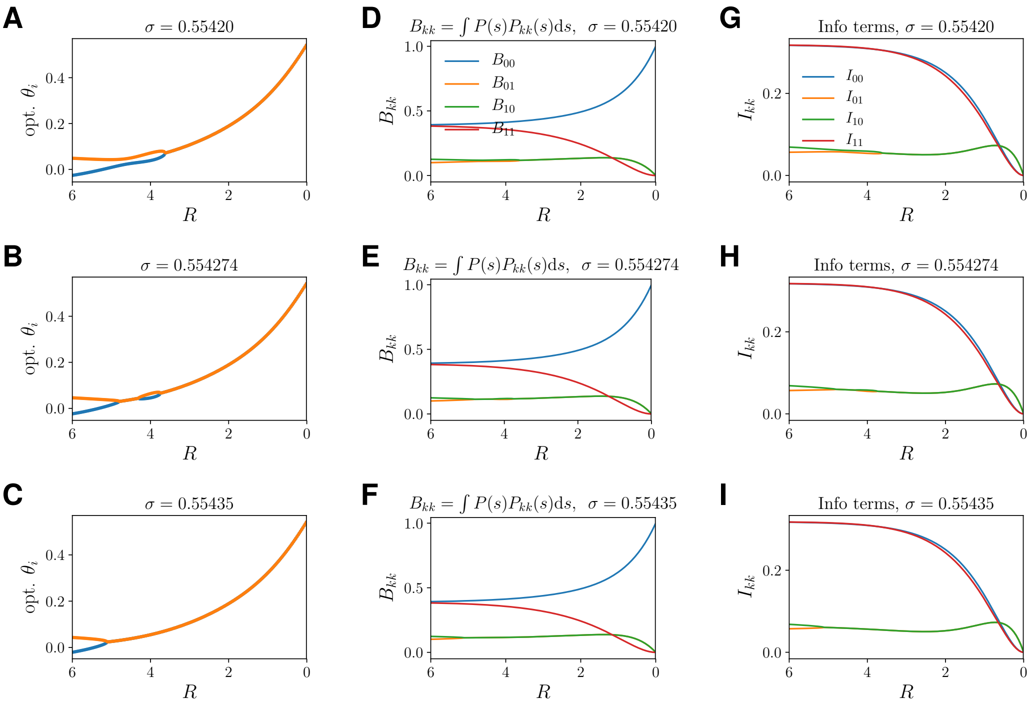


Figure 3.15: No poles or other numerical instabilities are visible in the marginalized output probabilities or the components of the mutual information in the case of non-monotonous threshold behavior. A-C. Optimal thresholds for three different values of input noise σ (one value where the unexpected non-monotonicity of diversity occurs (B), as well as one value below and above (A and C, respectively)). **D-F.** For the respective input noise values and corresponding optimal thresholds all possible values of the marginalized probability distribution of the response, B_{kk} (Eq. 3.34) are shown. **G-I.** The respective parts of the total mutual information (Eq. 3.32). None of the values in (D-I) shows any infinite behavior for $R \gtrsim 1$.

have finite values except for $R \to 0$ (Fig. 3.15G-I). To avoid numerical instabilities due to

$$\lim_{x,y \to 0} y \log(x) = 0, \tag{3.35}$$

two different approaches were used to numerically calculate

$$P_{kk}(s) \log_2 \left(\frac{P_{kk}(s)}{B_{kk}} \right) = P_{kk}(s) \log_2 \left(P_{kk}(s) \right) - P_{kk}(s) \log_2 (B_{kk}) \tag{3.36}$$

in the cases of $P_{kk}(s)$ and/or B_{kk} becoming very small: The first approach is to add machine epsilon $\epsilon \approx 2.2 \cdot 10^{-16}$ to all $P_{kk}(s)$ and B_{kk}. Then, Eq. 3.36 takes a value basically in the order of ϵ for $P_{kk}(s) \to 0$ (since $\log_2(\epsilon) = -52$). The second approach is to set $P_{kk}(s) = 1$ (since $\log_2(1) = 0$) for all[16] s where $P_{kk}(s)$ is smaller

[16] Note that I discretized s in the numerical calculations.

than a very small threshold value ϑ):

$$P_{kk}(s)) = \begin{cases} 1 & \text{for } P_{kk}(s) < \vartheta \\ P_{kk}(s) & \text{for } P_{kk}(s) \geq \vartheta \end{cases} \tag{3.37}$$

Both $\vartheta = 10^{-12}$ and 10^{-14} were used, but there was no difference between using these two values. Additionally, I set

$$\log_2(B_{kk}) = \begin{cases} 0 & \text{for } B_{kk} < \vartheta \\ \log_2(B_{kk}) & \text{for } B_{kk} \geq \vartheta \end{cases}. \tag{3.38}$$

Both of the two approaches lead to the same results of optimal thresholds for all noise combinations (not shown) and thus indicate that numerical instabilities due to Equation 3.35 were avoided since otherwise the two different approaches or the different values of ϑ would lead to significantly distinct results.

Nevertheless, I also want to point out that there is no pole-like behavior in $P_{kk}(s)$, as all values go to 0 or 1 for $s \to \pm\infty$ (data not shown). Furthermore, the $P_{kk}(s)/B_{kk}$ show no suspicious behavior: since this term depends on the stimulus it can only be plotted in a comprehensible way for one combination of noise values at a time. As input noise values I chose again $\sigma \in \{0.5542, 0.554274, 0.55435\}$ and as output noise values I chose four values at and around the ribbon, i.e. $R \in \{3.0, 4.0, 4.6, 6.0\}$. For all four combinations of $\{k_i\}$ there are no visible irregularities of $P_{kk}(s)/B_{kk}$ for any of the twelve noise combinations (Fig. 3.16).

In summary, I can rule out all numerical instabilities I could have thought of and thus conclude that the surprising non-monotonic behavior of threshold diversity described in this section is not caused by numerical implementations but is the true behavior.

3.5.3 Relationship to ubiquitous non-monotonic threshold differences

I have spent a considerable amount on time searching if similar results of non-monotonicity in efficient coding or other optimization problems exist in the literature. For the non-monotonic behavior of optimal threshold diversity, I only found one similar result, which has been observed when optimizing Fisher information – a different, local measure for information – in a population of bell-shaped tuning curves in a model of optimal coding of interaural time differences in the auditory brain stem [84]. There, the optimal number of distinct thresholds is 2-3-2 with increasing encoding precision, however, the authors did not comment on this phenomenon.

When looking at optimal threshold *differences*, $\theta_i - \theta_j$, with changing input noise in my work, these do appear for almost the whole range of output noise (Figs. 3.17C,E;

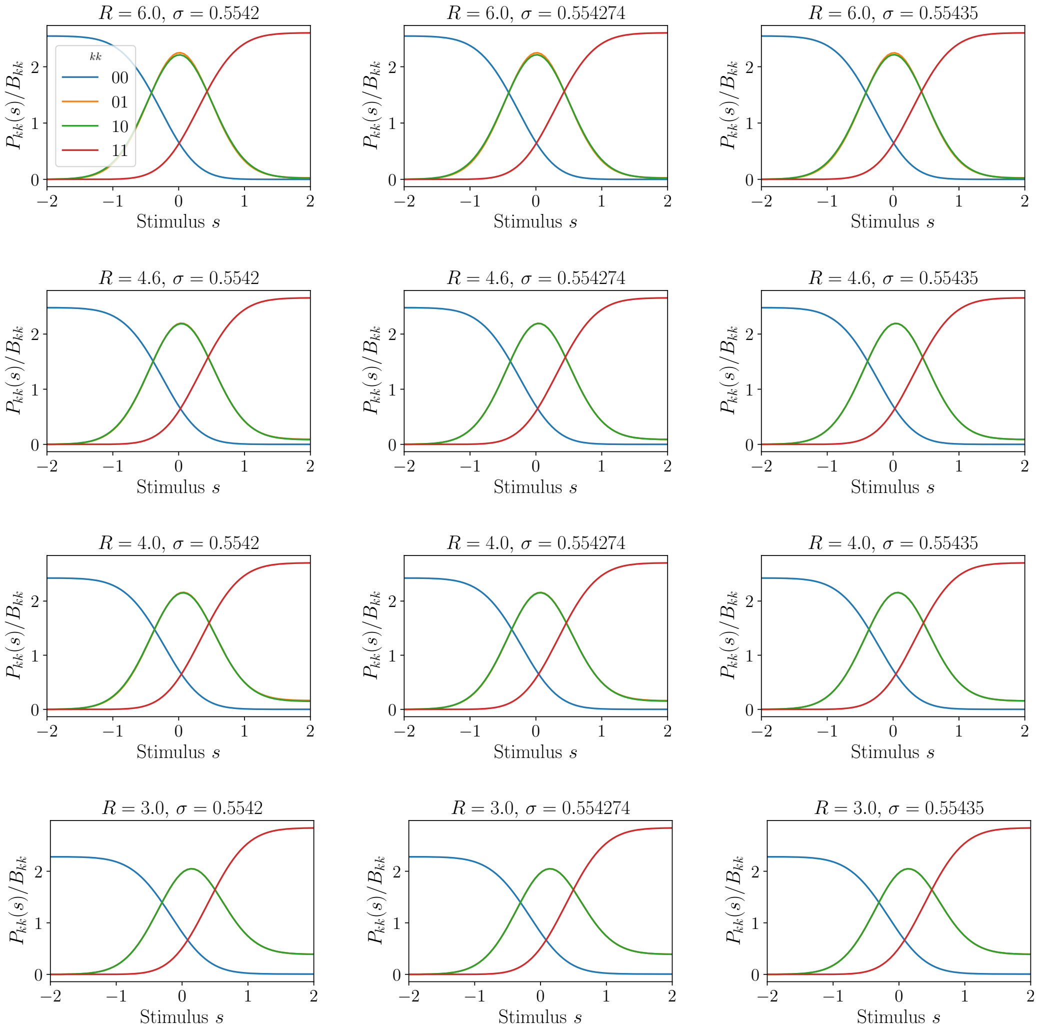


Figure 3.16: No poles or other numerical instabilities are visible in in the logarithmic terms of the mutual information. For $N = 2$, $P_{kk}(s)/B_{kk}$ (see Eqs. 3.33 and 3.34) are shown for three different values of input noise σ (along columns) and four different values of output noise R (along rows) for the corresponding optimal thresholds shown in Fig. 3.15A-C. The middle column shows terms with the input noise value that leads to non-monotonic behavior with output noise.

3.5C,E; 3.4C,E), while the non-monotonicity of threshold differences with changing output noise does only appear at a very small range of input noise (Figs. 3.17D,F,G). These non-monotonic threshold differences with input noise have been found before [62,85] and do also occur when using a mean rate constraint [29] (see Fig. 3.24 in Sec. 3.6.3D), but have so far not been commented on. Concluding, the non-monotonic behavior of threshold differences is not an uncommon phenomenon and can be seen as a less peculiar version of the non-monotonic behavior of threshold

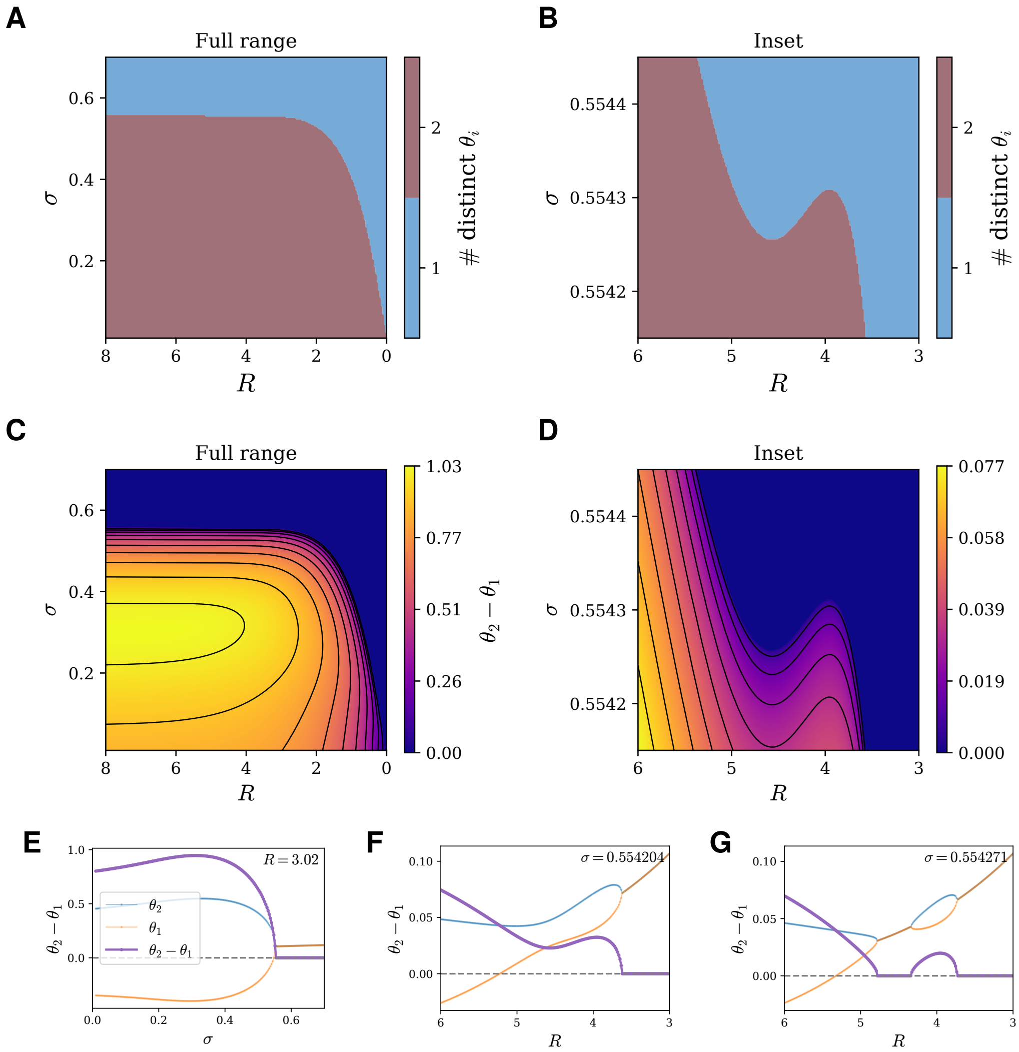


Figure 3.17: Threshold differences change non-monotonously with noise levels. For the independent-coding channel and $N = 2$ neurons. **A.** The optimal threshold diversity for a wide range of the noise space (input noise σ and output noise R). **B.** Inset of A in the range of noise space at which the non-monotonicity of optimal diversity with R occurs. **C, D.** The optimal threshold differences $\theta_2 - \theta_1$ of (A,B) are color-coded. **E-G.** Slices through C, D when keeping one noise level fixed shows that optimal threshold differences change non-monotonously with the other noise level varied. Note that E and F show a non-monotonicity of the threshold differences with input noise and output noise respectively, but both show no non-monotonicity of the number of distinct optimal thresholds.

diversity. As with the whole topic of non-monotonic threshold behavior, there is still a lack of explanations for these occurrences and their relationships.

3.6 Using different noise models

All results so far have been obtained by using a Gaussian normal distribution as input noise distribution and Poisson noise as output noise. In this section, I will investigate if different noise models lead to qualitatively different results, in particular, if the non-monotonic threshold diversities are restricted to the normally distributed input noise. In addition, I will treat the case where the input noise fed to each neuron is not equal but distinct for each neuron. Since this latter case becomes involved quickly for larger populations I will restrict my analysis to the two-neuron setup. All of these investigations will be done for the independent-coding channel since for this channel type the non-monotonic behavior occurs and the computational load is small.

3.6.1 Effect of the input noise model

Here, I use two different classes of noise distributions, namely the generalized normal distribution and a distribution related to the logistic function.

3.6.1.1 Generalized normal distribution

Remark: The methods of this subsection (Eqs. 3.40, 3.41, 3.43-3.45) have been described briefly in [1], and two figures (Fig. 3.18C,G) have been published in the supplementary material of [1] and were also briefly mentioned in the main text. Here, I describe the methods in more detail and present and describe the results in a greater context and a more detailed manner.

The Gaussian distribution has just two parameters, μ and σ, which define the mean and the standard deviation of the distribution, respectively. However, distributions have more properties than just these two, for example how asymmetric a distribution is around its mean, or how heavy the tails of a distribution are. The asymmetry is called *skewness*, and the heaviness of the tails is called *kurtosis* [103]. The generalized normal distribution (GND) generalizes the Gaussian distribution [104]. It is still symmetric around the mean, but it has an additional parameter which allows to vary the kurtosis. The kurtosis of a random variable z with mean μ is defined as [103]

$$\mathrm{Kurt}[z] = \langle (z-\mu)^4 \rangle. \tag{3.39}$$

All Gaussian distributions have a kurtosis of three. In analogy to the mean and the variance being called the first and second moments, respectively, the kurtosis is called the fourth moment.[17] The GND is parameterized by the mean (which will be set to zero as with the Gaussian noise) and two additional parameters, α and β,

[17] Similarly, the skewness is defined as $\langle (z-\mu)^3 \rangle$ and is thus called the third moment.

and is given as [104]

$$p_{\mathrm{GND}}(z) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|z|/\alpha)^{\beta}} \tag{3.40}$$

where $\Gamma(x)$ is the gamma function given as

$$\Gamma(x) = \int_0^{\infty} h^{x-1} e^{-h} \mathrm{d}h. \tag{3.41}$$

The parameter β controls the kurtosis such that

$$\mathrm{Kurt}[z] = \frac{\Gamma(1/\beta)\Gamma(5/\beta)}{\Gamma^2(3/\beta)} \tag{3.42}$$

and the parameter α is connected to the standard deviation σ via

$$\alpha = \sigma\sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}}\ . \tag{3.43}$$

As before, the standard deviation of the input noise distribution, σ, compared to the standard deviation of the stimulus distribution (the latter is set to $\sigma_s = 1$) determines the level of the input noise. The effective tuning curve of each neuron (Eq. 3.14) with the GND as input noise is given as

$$H_i(s) = \frac{1}{2} - \mathrm{sign}(\theta_i - s)\frac{\gamma\left(1/\beta, \left(\frac{|\theta_i - s|}{\alpha}\right)^{\beta}\right)}{2\Gamma(1/\beta)} \tag{3.44}$$

where $\gamma(x, y)$ is the lower incomplete gamma function defined as

$$\gamma(x, y) = \int_0^{y} t^{x-1} e^{-t} \mathrm{d}t\,. \tag{3.45}$$

The GND relaxes to well-known distributions, namely to the Laplace for $\beta = 1$, to the classic Gaussian distribution for $\beta = 2$, and the uniform distribution for $\beta \to \infty$. The distributions with different β are plotted in Fig. 3.18A. Remember that the effective tuning curve is given as the cumulative distribution function of the dichotomized noise distribution (Eq. 3.12 and see Fig. 3.1C for visualization). Figure 3.18B shows the effective tuning curves for the different noise distributions. In the following, I will use the GND as the input noise distribution, and by varying its parameter β, I can investigate how the kurtosis of the noise distribution influences the optimal thresholds of the independent-coding channel.

As with the classical Gaussian noise, the results with a population of three neurons is that for low noise having three distinct thresholds is optimal, while for high noise having equal thresholds is optimal, with an intermediate range of having

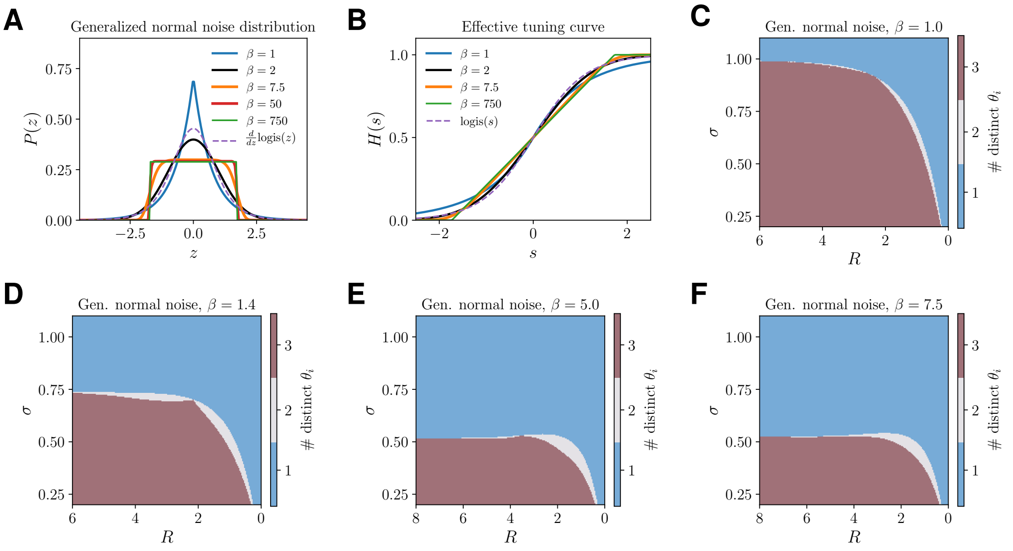


Figure 3.18: Effect of the kurtosis of the input noise distribution on the optial number of distinct thresholds. The different input noise distributions used and their effects on optimal threshold diversity across the noise space (R, σ are shown for a population of $N = 3$ neurons. **A.** Different noise distributions $P(z)$ obtained by using the generalized normal distribution (Eq. 3.40) with different values of the kurtosis parameter β. For comparison, in dashed is shown the logistic distribution (Eq. 3.48) which is the derivative of the logistic tuning curve. The variance is $\sigma^2 = 1$ for all distributions. **B.** Effective tuning curves $H(s)$ (see Eq. 3.12) for the noise distributions shown in (A) with $\theta = 0$. The effective tuning curves are given as the cumulative distribution functions of the noise distributions. The dashed line shows the logistic tuning curve (Eq. 3.46). **C-F.** Kurtosis of the noise distribution influences the non-monotonicity of optimal number of distinct thresholds. **C.** For $\beta = 1$ (meaning the noise distribution follows a Laplace), no non-monotonicity of optimal threshold diversity occurs. **D, E.** For β in the range of $\beta \approx [1.4, 5]$ there is a non-monotonicity of optimal threshold diversity for certain input noise levels σ due to threshold switching (compare to Fig. 3.4 where $\beta = 2$). **F.** For larger β – the noise distribution resembles more and more a uniform one – the threshold switching disappears while the non-monotonicity due to ribbons become more pronounced. Panels C and F are from [1].

two distinct optimal thresholds (Fig. 3.18C-F, compare to previous results with $\beta = 2$, Fig. 3.4A). For $\beta = 1$ (equivalent to Laplace noise[18]), there exists no non-monotonicity of optimal threshold diversity. The range of noise combinations for which a partial diversity – two distinct optimal thresholds – is optimal is very small. For a certain range of the kurtosis parameter, namely $\beta \approx [1.4, 5]$, there exists a non-monotonic behavior of optimal threshold diversity (Fig. 3.18D,E) due to threshold switching (not shown). This is in accordance with the previous result of the classical Gaussian distribution with $\beta = 2$ (Fig. 3.12B). For $\beta = 7.5$, the

[18] I also implemented Laplace noise using the classical expression for its cumulative distribution function and it conformed with the implementation using the generalized normal distribution for $\beta = 1$ (not shown).

non-monotonicity due to threshold switching also does not exist, but the ribbon non-monotonicity (shown in Fig. 3.12C for $\beta = 2$) is present and even exists for a larger range of input noise strength σ (the bump of the beige color into the blue color is larger in Fig. 3.18F compared to Fig. 3.4A).

I also used the GND as the input noise distribution for a population of four neurons (Fig. 3.19). For relatively high kurtosis ($\beta \lesssim 2.5$) the results are qualitatively similar to the case of three neurons: optimal thresholds bifurcate in subsequent transitions, i.e. there is a gradual change of optimal diversities. However, the range where the optimal number of distinct thresholds is three, is again quite small for Laplace noise, i.e. very high kurtosis ($\beta = 1$, Fig. 3.19A). For $\beta = 1$, there is also again no non-monotonic behavior for any noise value, while for $\beta \gtrsim 2$ the qualitative structure of non-monotonic behavior seen with the Gaussian noise is preserved (Fig. 3.19B,C). For $\beta = 5$ however, qualitatively new behavior occurs: contrary to $\beta = 2$, where only ribbon non-monotonicities are found (Fig. 3.12D,E) and where all threshold switching[19] are discontinuous and thus do not effect optimal diversity (Fig. 3.5B, arrow), for $\beta = 5$, now there also exists a non-monotonicity due to continuous threshold switching (Fig. 3.19E). Furthermore there exist ribbon non-monotonicities with respect to input noise σ (Fig. 3.19F), which have not been observed for $\beta = 2$.

Since the GND only becomes uniform for $\beta \to \infty$, I also implemented the uniform noise explicitly and compared it for high values of β. The explicit implementation of the uniform noise needs a high number of integration steps to be precise, thus I have here only done it for $N = 2$ neurons: For $\beta = 20$, the shape of optimal thresholds still differs somewhat (Fig. 3.20B), while for $\beta = 50$ the thresholds for most noise values are very similar and only differ for noise values near critical noise values (Fig. 3.20A).

3.6.1.2 Logistic distribution

In addition to effective tuning curves used so far which are described by being the cumulative distribution function of a (generalized) normal distribution, I also looked at tuning curves described by the logistic function [105]

$$\text{logis}(s) = \frac{1}{1 + e^{(\theta - s)/a}} \tag{3.46}$$

with variance

$$\sigma^2 = \frac{a^2\pi^2}{3}. \tag{3.47}$$

[19] The discontinuous transition of optimal thresholds from $\theta_1 = \theta_2 \neq \theta_3 = \theta_4$ to $\theta_1 = \theta_2 = \theta_3 \neq \theta_4$.

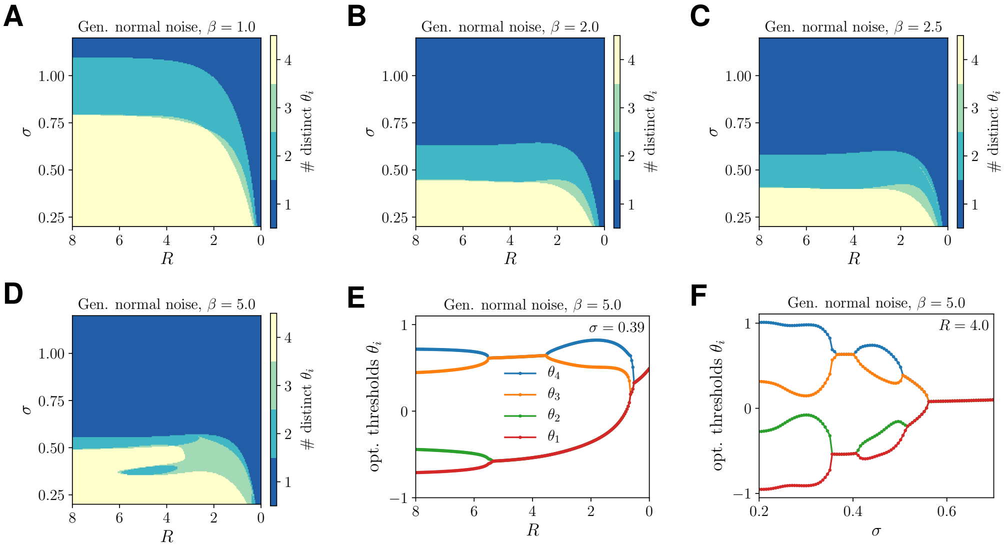


Figure 3.19: The effect of having a generalized normal noise distribution with four neurons and different kurtosis of the input noise distribution. A-D. The optimal number of distinct thresholds. **A.** There is no non-monotonicity of optimal threshold diversity for input noise distributions with high kurtosis ($\beta = 1$, i.e. for Laplace noise). **B,C.** For input noise distributions with kurtosis of a normal distribution ($\beta = 2$, B), or smaller kurtosis ($\beta = 2.5$, C) there occurs a non-monotonicity of optimal threshold diversity. **D.** For quite low kurtosis ($\beta = 5$), there are very pronounced non-monotonicities over a large area of the noise space. **E.** For $\beta = 5$, the optimal threshold values are shown in dependence of output noise while the input noise is fixed to $\sigma = 0.39$. Unlike for $N = 4$ with Gaussian input noise, one sees continuous threshold switching along the output noise axis. **F.** For $\beta = 5$, the optimal threshold values are shown in dependence of input noise while the output noise is fixed to $R = 4$. Unlike for Gaussian input noise, one sees threshold ribbons along the input noise axis.

The corresponding noise distribution is its derivative, namely

$$p_l(z) = \frac{e^{-z/a}}{a(1 + e^{-z/a})^2} \tag{3.48}$$

and is called the logistic distribution [105]. Both $p_l(z)$ and logis(s) are shown in Figure 3.18A and B, respectively. As with the GND as input noise distribution, the qualitative results of optimal thresholds remain the same (Fig. 3.21A). As such, the non-monotonic behavior of optimal threshold diversity through threshold switching for $N = 3$ also exists (Fig. 3.21B). However, threshold ribbons do not occur when using the logistic distribution as input noise distribution. The reason why I considered this noise distribution is that in principle it could be possible to find an analytical expression, $f(\sigma_c, R_c)$, for the combinations of critical noise values σ_c, R_c, at which the bifurcations occur; i.e. to find a function that describes the curves in the σ-R-plane at which the bifurcations occur. With $N = 2$ neurons, there is just a single curve, separating the noise space where the optimal thresholds are equal

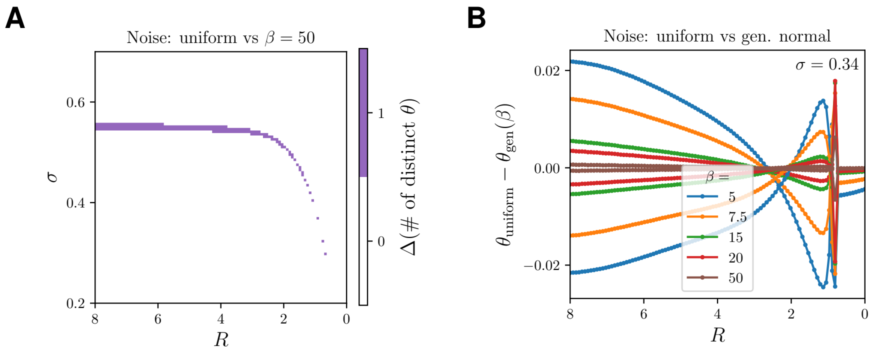


Figure 3.20: Comparing uniform noise with generalized normal noise with very low kurtosis, i.e. large β. A population of two neurons and the independent-coding channel is used. **A.** The difference in number of distinct thresholds between using real uniform noise and generalized normal noise with $\beta = 50$. For such a large β, there is only seen a small difference at critical noise values (purple region). **B.** Differences of optimal thresholds between uniform noise (optimal thresholds θ_{uniform}) and generalized normal noise (optimal thresholds θ_{gen}) for different parameters of β.

and the noise space where optimal thresholds are distinct. For that case, there are four unknowns $(\theta_1, \theta_2, \sigma_c, R_c)$ while there are three equations which are fulfilled at the bifurcations:

$$(I) \quad \frac{\partial I_m}{\partial \theta_1} = 0 \tag{3.49}$$

$$(II) \quad \frac{\partial I_m}{\partial \theta_2} = 0 \tag{3.50}$$

$$(III) \quad \lambda_2 = 0 \tag{3.51}$$

where λ_2 is the smaller eigenvalue of the Hessian at θ_1, θ_2 (see Sec. 4.3 for details and why Eq. III is true). This set of equations would allow me to find a relationship between R_c and σ_c for two neurons.[20] It became apparent that at least Eq. I and II cannot be solved analytically in the case of having the classical Gaussian as the input noise distribution $p(z)$: the thresholds influence the mutual information through the effective tuning curves $H_i(s)$ (see Eqs. 3.12 to 3.20):

$$H_i(s) = P(\nu_i = \nu_{\max}|s) = \int_{\theta_i - s}^{\infty} \mathrm{d}z \, p(z) \tag{3.52}$$

and when $p(z)$ is Gaussian then $H_i(s)$ is not analytically solvable. When $p(z)$ is Laplacian or uniform there have to be made extensive case distinctions – e.g. if θ_1, and/or θ_2 are larger/smaller than $\pm\sigma$ – for integration limits. Using the logistic distribution, $p_l(z)$, (Eq. 3.48) for $p(z)$, a closed and continuously differentiable form

[20] However, this approach would in principle also work for larger N, as more equations become available to use.

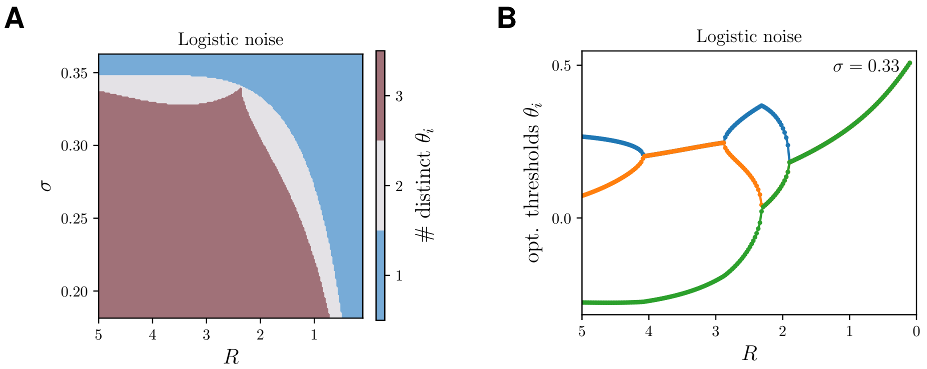


Figure 3.21: Using a logistic function as the effective tuning curve with a population of three neurons and the independent-coding channel. **A.** Optimal number of distinct thresholds color-coed in the input-output-noise space (σ, R). **B.** Optimal thresholds for varying output noise R and input noise fixed to $\sigma = 0.33$. As with Gaussian input noise there happens a threshold switching with R for a range of σ.

of $H_i(s)$ exists since logis(s) is by definition the integral of p_l. Thus

$$H_i(s) = \left.\frac{1}{1+e^{-z}}\right|_{\theta_i - s}^{\infty} = 1 - \frac{1}{1+e^{-(\theta_i - s)}} \quad . \tag{3.53}$$

For $N = 2$, the information consists of four terms ($k_i = \{0, 1\}$ for $i = 1, 2$). Using the notation as in Sec. 3.5.2.3, the first term of the mutual information is

$$I_{00} = \int_s P_s(s) Q_1(s) Q_2(s) \log_2 \left(\frac{Q_1(s) Q_2(s)}{\int_{s'} \mathrm{d}s' P_s(s') Q_1(s') Q_2(s')} \right) \mathrm{d}s. \tag{3.54}$$

Expressing $Q_i(s)$ in terms of $H_i(s)$ (see Eq. 3.18) one obtains

$$\begin{aligned} I_{00} = \int \frac{e^{\frac{-s}{2}}}{2(1+e^{\frac{-s}{2}})^2} \left[\left(1 - \frac{1}{1+e^{-(\theta_1 - s)/\sigma}}\right) + \frac{e^R}{1+e^{-(\theta_i - s)/\sigma}} \right] \\ \cdot \log_2 \left(\frac{\left(1 - \frac{1}{1+e^{-(\theta_1 - s)/\sigma}}\right) + \frac{e^R}{1+e^{-(\theta_i - s)/\sigma}}}{\int \frac{e^{\frac{-s'}{2}}}{2(1+e^{\frac{-s'}{2}})^2} \left[\left(1 - \frac{1}{1+e^{-(\theta_1 - s')/\sigma}}\right) + \frac{e^R}{1+e^{-(\theta_i - s')/\sigma}} \right] \mathrm{d}s'} \right) \mathrm{d}s \end{aligned} \tag{3.55}$$

To use equations $(I) - (III)$, one should eliminate θ_1 and θ_2. However, that seems very difficult, especially since Eq. 3.55 also has to be differentiated once – for (I, II) / Eqs. 3.49, 3.50 – or twice – for (III) / Eq. 3.51 – with respect to $\theta_{1,2}$. I tried to obtain expressions for (I, II) using Mathematica (Wolfram Research), however, the results after several hours of computations were expressions stretching over several pages, from which it was not feasible to eliminate θ_1 and θ_2 as they occurred all over the place. Nevertheless, I still wanted to outline a principle possibility of how to potentially calculate critical noise levels analytically.

In conclusion, I can say that the general result of having a series of bifurcations of optimal thresholds with varying either input or output noise also holds when having an input noise distribution different from Gaussian. Also the non-monotonic threshold behavior of threshold switching and threshold ribbons exist for non-Gaussian noise. The conclusion is, however, that non-monotonicities are more pronounced and occur in a wider parameter range for input noise distributions with smaller kurtosis – like the uniform distribution – while they do not occur at all for distributions with high kurtosis – like the Laplace distribution. The reason might be that the tails of the stimulus distribution get more attention with higher kurtosis of the input noise distribution. Thus there would be a greater overlap of tuning curves in the stimulus space where their probabilities of firing with maximum firing rate is significantly different from zero or one. However, so far I have not found an intuitive explanation why this difference in kurtosis should cause a difference in non-monotonic behavior.

3.6.2 Distinct input noise levels

Often in neural populations, tuning curves differ both in their thresholds and in their steepnesses [52,56]. By using binary tuning curves with distinct levels of additive input noise, one obtains effective tuning curves with distinct values of steepness. Thus, I introduced a noise difference $\Delta\sigma$ such that $\sigma_1 = \sigma_2 - \Delta\sigma$. Apart from different σ_i the setup is the same as before, in specific, I use the classical Gaussian input noise again and varied both the input noise and output noise.

First, I found that there exist at least one local maximum, thus I used a grid of initial thresholds $\vec{\theta_0} = \{\theta_{0,1}, \theta_{0,2}\}$ with 20 steps for each $\theta_{0,i}$. Using that grid and plotting the information landscape, I found that independent of output noise level, R, there is only one local maximum in addition to the global maximum (Fig. 3.22A,B). I found that this local maximum can be avoided reliably for all R by using five steps for each $\theta_{0,i}$.

The main result is that for low output noise it is optimal that the neuron with steeper effective tuning curve – i.e. with less input noise – has the higher threshold (Fig. 3.22A,C), while for high output noise the opposite is optimal – i.e. the neuron with shallower tuning curve having the higher threshold (Fig. 3.22B,D). However, the two possibilities have only a small difference in information transmission: For small output noise the relative difference is of the order of 10^{-5} bits, while for high output noise it is on the order of 10^{-3} bits (not shown).

Then I extended the investigations to using arrays of different input noise differences, $\Delta\sigma$, and input noise values of neuron two, σ_2, as well as an array of output noise values R. The input noise value of neuron one is automatically given as

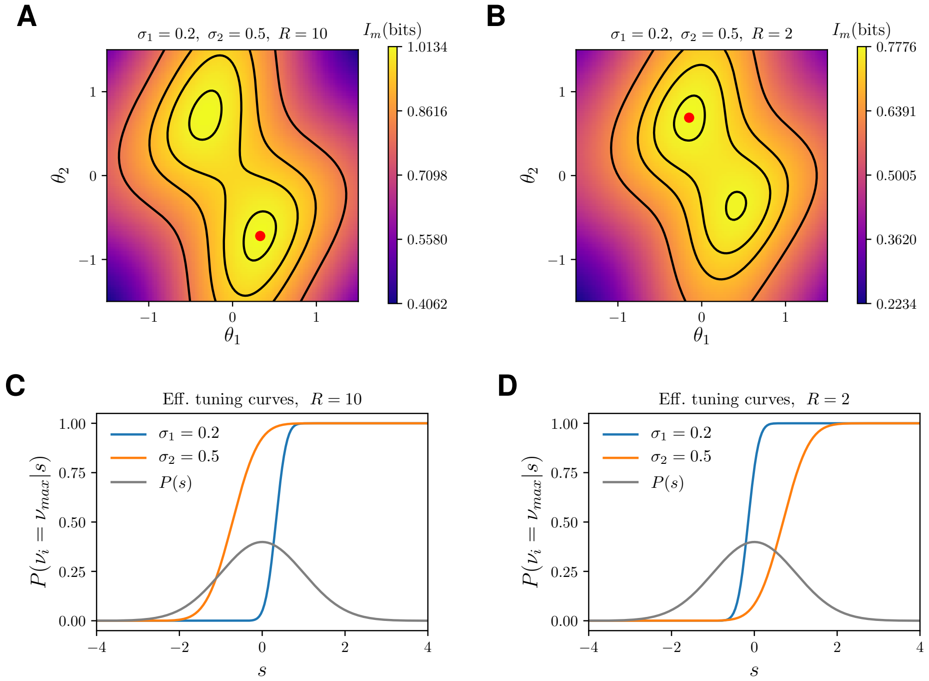


Figure 3.22: Having distinct input noise levels for each neuron introduces an asymmetry of the information landscape. Independent-coding channel with two neurons and two distinct input noise levels, σ_1 and σ_2. **A, B.** The information landscape $I_m(\theta_1, \theta_2)$ has two maxima, of which only one is the global one (denoted by the red dot). When changing output noise levels from low (A: $R = 10$) to high (B: $R = 2$) the global maximum swaps along the diagonal of the thresholds space. **C, D.** Respective optimal tuning curves of the two neurons (orange and blue) from (A,B). The stimulus distribution is shown for comparison in gray. The swap of the global maximum means that while for low output noise it is optimal that the neuron with a steeper tuning curve (with smaller σ, blue) has the larger threshold and is thus positioned to the right of the other neuron on the stimulus axis (C). On the contrary, for high output noise it is optimal that the neuron with the shallower tuning curve (larger σ, orange) has the larger threshold and is positioned right of the other neuron (D).

$\sigma_1 = \sigma_2 - \Delta\sigma$ for each combination of $\Delta\sigma$ and σ_2. Figure 3.23A,B shows how the optimal threshold difference $\Delta\theta$ depends on $\Delta\sigma$ for two values of R. In accordance with the results shown above, there is a sign switch of optimal $\Delta\theta$ with changing output noise R (compare Fig. 3.23A where $R = 10.0$ with Fig. 3.23B where $R = 2.3$).[21] From this result, one could postulate that there exists a critical value of R somewhere between 2.3 and 10.0, for which the sign switch of optimal $\Delta\theta$

[21] Note that the optimal $\Delta\theta$ are not symmetric around $\Delta\sigma = 0$ since $\sigma_1 = \sigma_2 - \Delta\sigma$ for each σ_2 so that the total noise $\sigma_1 + \sigma_2 = -\Delta\sigma$ varies with $\Delta\sigma$. Therefore, for example, for $\Delta\sigma > 0.1$, all $\Delta\theta$ are $\neq 0$ since the total noise is so small that the optimal solution is to have two distinct thresholds for all σ_2 shown.

happens.[22] However, $\Delta\theta$ does not happen for all σ_2 at the same R: For $R = 7.8$, for example, there is only a sign switch for low values of σ_2 (Fig. 3.23C,D); while a sign switch for larger σ_2 only occurs at $R > 7.8$. This means that both output and input noise level influence for which R when these sign switches of $\Delta\theta$ occur.

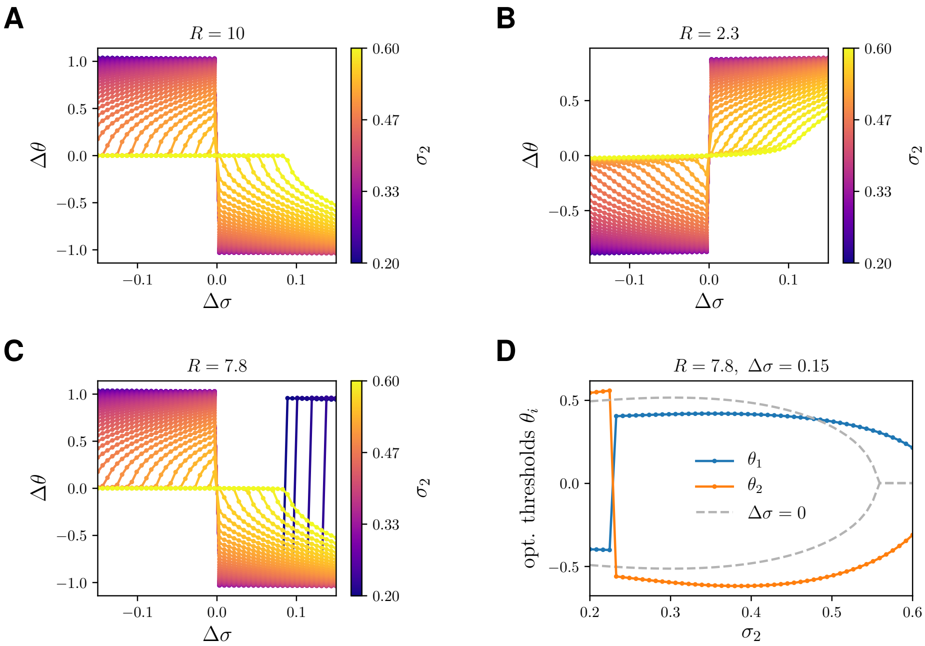


Figure 3.23: Optimal threshold differences $\Delta\theta$ depend on input noise differences $\Delta\sigma$. A, B. Between low ($R = 10.0$, A) and high ($R = 2.3$, B) output noise, $\Delta\theta$ switches signs. **C, D.** For an intermediate output noise level ($R = 7.8$), it is shown how the switch between the two maxima takes place (here at $\sigma_2 \approx 0.25$ and $\Delta\sigma \approx 0.1$). The occurrence of the switching depends on R, $\Delta\sigma$, and σ_2.

3.6.3 Distinct input noise values with a mean rate constraint

A setup with two neurons and distinct input noise levels has already been investigated by Kastner et al. [29]. Contrary to the one described above, they have included a mean rate constraint and no output noise. Until now I have given threshold values in the original stimulus space s with $p(s) \sim \mathcal{N}(0, 1)$. For some setups, e.g. the one in this section, the shape of the stimulus distribution has no effect on optimal thresholds. To compare thresholds across any stimulus distribution, it is beneficial to use the cumulative distribution function of the stimulus distribution, given as

$$s' = \int_{-\infty}^{s'} p_s(s)\mathrm{d}s \ . \tag{3.56}$$

[22] Note that I chose these seemingly arbitrary values of 2.3, 7.8 and 10.0 for R just because I ran a grid of 10 values between 0.1 and 10.

Kastner et al. introduced a mean rate constraint through $\langle r\rangle$:

$$\langle r\rangle = \theta'_1 + \theta'_2 \ , \tag{3.57}$$

where θ'_i are thresholds in the cumulative stimulus space. Thus, $0 < \langle r\rangle < 2$, with results of $0 < \langle r\rangle < 1$ being symmetric to those of $1 < \langle r\rangle < 2$, which is why I only did calculations for $0 < \langle r\rangle < 1$. Again, I set $\sigma_1 = \sigma_2 - \Delta\sigma$. Only one threshold is optimized for each combination of $\{\langle r\rangle, \Delta\sigma, \sigma_2\}$ and the other threshold is obtained by fulfilling Eq. 3.57. The goal is first to reproduce a finding of Kastner et al., namely that distinct input noise levels ($\sigma_1 \neq \sigma_2$) introduce an asymmetry among the thresholds, such that $\Delta\theta < 0$ if $\Delta\sigma < 0$, and $\Delta\theta > 0$ if $\Delta\sigma > 0$ [29]. The results are shown in Fig. 3.24. As in the previous paragraph (with output noise and no mean rate constraint), there are sign switches of optimal $\Delta\theta$ even without a sign switch of $\Delta\sigma$ when varying σ_1 while keeping σ_2 fixed (Fig. 3.24A,C).[23] This result is in contrast to Figure 2A of Kastner et al.'s work [29] since they do not observe sign switches of $\Delta\theta$ without sign switches of $\Delta\sigma$. They even explicitly state that $\Delta\theta < 0$ if $\Delta\sigma < 0$, and $\Delta\theta > 0$ if $\Delta\sigma > 0$ for any $\sigma_2 \neq \sigma_1$. Possible reasons could be that (1) they did not take into consideration the switch from the global to the local maximum (they do not mention local maxima anywhere in their publication), or (2) they chose large σ_i only, for which the sign switches do not appear (Fig. 3.24B) (it is not clear which values of σ_i they use as they only denote the values of $\Delta\sigma$). Another result is that using the mean rate constraint of Kastner et al. qualitatively changes the behavior of optimal thresholds: no bifurcations appear for any input noise level combinations (σ_1, σ_2), as long as $\sigma_1 \neq \sigma_2$ (Fig. 3.24C,D). The results so far show that more research has to be done to understand where the differences between my and Kastner et al.'s results come from. Note that the absence of threshold bifurcations seen in Fig. 3.24C,D is in accordance with their results.

The setup by Kastner et al. can be generalized to also incorporate output noise: instead of having a firing probability of one when the input is above the threshold, one sets it to $r_i = 1 - e^{-R_i}$ (see Eq. 3.15), again with $R_i = \nu_{max,i}\Delta T$. Then, the overall mean rate is

$$\langle r\rangle = r_1\theta'_1 + r_2\theta'_2. \tag{3.58}$$

Two additional parameters have to be chosen in this case: r_1, and r_2. Again, only one threshold is optimized for each combination of$(\langle r\rangle, r_1, r_2, \sigma_1, \sigma_2)$ and the other threshold is obtained by fulfilling Eq. 3.58. To not make things too complicated, one can assume that the two neurons have the same maximum firing rates (i.e. the same output noise levels), thus $r_1 = r_2$. Still, the whole story would become quite involved since optimizations for combinations of four parameters are performed. To not get lost in details in this work, these studies remain to be done in the future.

[23] Note, that to reproduce Kastner at al.'s results the mean rate $\langle r\rangle$ is now color-coded and σ_2 is fixed in Figure 3.24A,B.

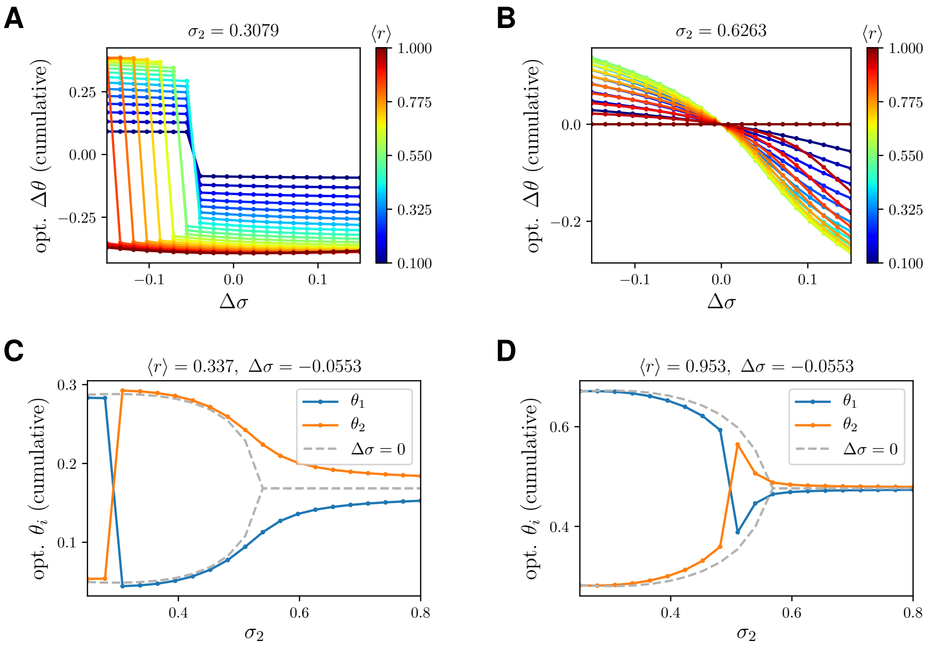


Figure 3.24: Threshold differences for only input noise with a mean rate constraint according to Kastner et al. [29]. Optimal thresholds differences $\Delta\theta$ depend on input noise differences $\Delta\sigma$ for different mean rates $\langle r \rangle$ (color coded) and input noise of neuron two, σ_2. **A.** For small σ_2, sign switches of $\Delta\theta$ occur without sign switches of $\Delta\sigma$. **B.** For large σ_2, sign switches of $\Delta\theta$ occur only at sign switches of $\Delta\sigma$. **C,D.** For a fixed value of the mean rate, $\langle r \rangle$, and a fixed value of input noise differences, $\Delta\sigma$, optimal thresholds are shown in dependence of input noise σ_2. As in the previous two figures (Fig. 3.22 and 3.23C,D) the discontinuous switching of the optimal thresholds corresponds to a switch of the global maximum along the diagonal of the (θ_1, θ_2)-space.

3.6.4 Effect of the output noise model

Now, I will investigate the effects of the output noise model optimal thresholds. The focus is not primarily on describing how the previous results change when varying the output noise model, but instead on giving an understanding about how a probabilistic spike generation process – which necessarily implies a discrete, non-zero output variable since spikes can only be non-zero integers – affects optimal thresholds. Therefore and for reduced computational load, I will now remove input noise from the system, i.e. I will have truly binary tuning curves again.

That way, I also hope to find explanations for two peculiarities regarding output noise: first, there is the phenomenon that output noise shifts the optimal thresholds away from the stimulus mean towards higher values of the stimulus space. The consequence is that the average of the optimal thresholds is always larger than the stimulus mean, except in the limit of no output noise, $R \to \infty$ (see how optimal thresholds are on average considerably larger than zero in Fig. 3.4 and 3.5, especially for small R). This phenomenon is especially pronounced for $R \to 0$, where it has been analytically calculated that optimal thresholds are at $1 - 1/e \approx 0.63$ in the cumulative stimulus space [106] (note that the stimulus mean is at 0.5 in the cumulative space). Second, with the independent-coding channel and only output noise, optimal thresholds are distinct for all finite R, i.e. the bifurcation happens at $R \to 0$ (see Figs. 3.4D, 3.5D, and 3.6C). Before demonstrating the effects of output noise models different from Poisson I will show how spontaneous firing rates affect optimal thresholds. From now on I will denote threshold values in the cumulative stimulus space since it has been shown that without input noise the shape of the stimulus distribution does not matter [86, 106].

3.6.4.1 Including spontaneous firing rates

The binary tuning curves I have used so far have a zero firing rate level, i.e. the firing rate is zero below the threshold. The Poisson distribution is deterministic for zero rate, so that $P(k_i = 0|\nu_i = 0) = 1$ and thus the noise entropy is zero for $\nu_i = 0$. This property might be the cause for the two peculiar phenomena of shifted thresholds towards the right of the stimulus mean and having distinct threshold for all $R > 0$. Here, I investigated the case of a binary tuning curve with both firing rates being larger zero, such that the tuning curve becomes

$$\nu_i(s) = \nu_0 + (\nu_{\max} - \nu_0)\Theta(\theta_i - s) \ . \tag{3.59}$$

As with the maximum rate, $\nu_{\max}$, the interesting size is the expected spike count in the coding window, $R_0 = \nu_0 \Delta T$, which can be directly compared to $R = \nu_{\max}\Delta T$. Since it is not possible anymore to map all spike counts larger zero to the high-firing rate level, analytical calculations were replaced by numerical ones (summing over all spike counts with significantly large probability, and finding the optimal thresholds using the SLSQP algorithm[24]). For simplicity and to reduce the computational load, I did calculations with a population size of only $N = 2$ neurons. However, as in Figure 3.4D, the two thresholds bifurcate at $R \to 0$ in all cases, three neurons would not provide more insights anyway.

As expected, larger spontaneous rates decreased the information transmission and thus act as another source of noise (Fig. 3.25). For very low spontaneous firing rate of $R_0 = 0.005$, the result resembles previous (analytical) results [106] with no spontaneous rate (Fig. 3.25A,B). For significant spontaneous firing rates, $R_0 = 0.05$ or $R_0 = 0.5$, the information is decreased (Fig. 3.25C,E) and the optimal thresholds are shifted towards the stimulus mean (Fig. 3.25D,F). As expected in the limit of vanishing noise, $R \to \infty$, both the information and the optimal thresholds resemble the case with no spontaneous rates: the maximized information reaches capacity

[24] *Sequential Least Squares Programming*, a local optimizer being able to handle constraints [96]. I performed optimization in the cumulative space with the constraint of $0 \leq \theta_2 \geq \theta_1 \leq 1$ what reduced the computational load.

at $\log_2(3)$ bits and the optimal thresholds become $\{\frac{1}{3},\frac{2}{3}\}$ in cumulative space. Regarding the optimal thresholds (Fig. 3.25D,F) there are two observations. First, the optimal thresholds are still distinct for all finite output noise values, what shows that this peculiar phenomenon is not caused alone by having zero spontaneous firing rates. Second, interestingly, the optimal thresholds for infinite output noise, $R \to 0$, are shifted towards the mean of the stimulus distribution (0.5 in cumulative stimulus space) with spontaneous rates. Brinkman et al. gave the following intuitive explanation for why optimal thresholds are shifted away from the mean with high output noise: the Poisson distribution has larger variance with larger firing rates and noise entropy is reduced by encoding the most probable stimuli with lower firing rates, i.e. by shifting thresholds towards higher values of the stimulus distribution [76]. The shifting back towards the stimulus mean is in accordance with this explanation since for smaller firing rates – i.e. R getting in the range of R_0 – the advantage of having less variance is reduced.

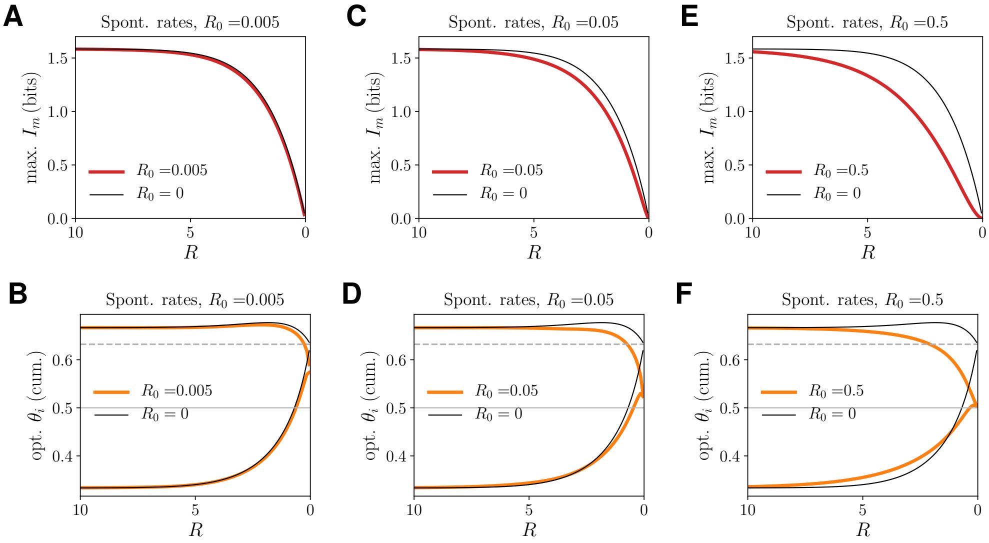


Figure 3.25: Spontaneous rates reduce information and shift optimal thresholds towards stimulus mean at high output noise. A. Maximized mutual information in the case of low spontaneous rate, $R_0 = 0.005$, is almost the same as without spontaneous rate, $R_0 = 0$. **B.** Optimal thresholds corresponding to A. Near $R \to 0$, there is a small shift from the theoretical value of $1 - 1/e$ for $R_0 = 0$ (dashed gray line) towards the stimulus mean of 0.5 (solid gray line). **C.** As A but with slightly larger spontaneous rate, $R_0 = 0.05$, and a slightly decreased information for intermediate output noise values. **D.** Optimal thresholds are significantly shifted towards the stimulus mean for $R \to 0$. **E.** As C but with $R_0 = 0.5$. **F.** As D but with $R_0 = 0.5$.

3.6.4.2 Changing the output noise model

After having shown that the presence of spontaneous rates cannot explain the phenomenon that optimal thresholds have distinct values for all finite output noise values (in the absence of input noise, Fig. 3.4D), the goal was to investigate if some specific properties of the Poisson noise are the cause for this phenomenon. A prominent property of the Poisson distribution is the fact that its variance is equal to the mean. Both are given by its rate parameter, λ:

$$\mu = \sigma^2 = \lambda \ . \tag{3.60}$$

The ratio between the variance and the mean is used to quantify the *dispersion* of a distribution and is called *Fano factor* [107]:

$$F = \frac{\sigma^2}{\mu} \ , \tag{3.61}$$

where the case of $F > 1$ is called over-dispersion and the case of $F < 1$ is called under-dispersion. Thus, my idea was to work with a discrete probability distribution for which the variance deviates from the mean. One can obtain such a distribution by modifying the traditional Poisson distribution in various ways [108], e.g. by using a compound Poisson distribution, by convolving a Poisson distribution with another distribution, or by just modifying its formula. Another possibility is to multiply the Poisson distribution with a function $F(k)$ which depends at least on k and optionally also on λ, and subsequently to ensure normalization. Note that in the latter case every function, $F(\lambda)$, that only depends on λ and not on k will lead to a normalization term of $e^{-\lambda}/F(\lambda)$.

The Delaport distribution

By convolving the Poisson distribution with a negative binomial distribution, $P_{NB}(k|\alpha, \beta)$, one obtains the *Delaporte distribution* [108]:

$$\begin{aligned} P_{\text{Delap}}(k|\lambda, \alpha, \beta) &= P_{\text{Poiss}}(k|\lambda) * P_{NB}(k|\alpha, \beta) && (3.62) \\ &= \sum_{i=0}^{k} P_{\text{Poiss}}(k-i|\alpha)\, P_{NB}(i|\alpha, \beta) && (3.63) \\ &= \sum_{i=0}^{k} e^{-\lambda}\ \frac{\lambda^{(k-i)}}{(k-i)!}\ \frac{\Gamma(\alpha+i)}{\Gamma(\alpha) i!} \beta^i (1-\beta)^\alpha && (3.64) \end{aligned}$$

where $\lambda, \alpha, \beta > 0$ and with

$$\Gamma(x) = \int_0^\infty h^{x-1} e^{-h} \mathrm{d}h. \tag{3.65}$$

The mean is given by

$$\mu = \lambda + \alpha\beta \tag{3.66}$$

and the variance as

$$\sigma^2 = \lambda + \alpha\beta(1+\beta) \ , \tag{3.67}$$

thus $\sigma^2 > \mu$ for all $\beta > 0$ and the Delaporte is always more dispersed than the original Poisson. For $\alpha, \beta \to 0$ the Delaporte converges to the Poisson distribution [108].

Conway-Maxwell-Poisson distribution

When multiplying the Poisson distribution with $F(k|\phi) = (k!)^{1-\phi}$ one obtaines the *Conway-Maxwell-Poisson* (CMP) distribution [109]:

$$P_{\mathrm{CMP}}(k|\lambda,\phi) = \frac{1}{Z(\lambda,\phi)} \frac{\lambda^k}{(k!)^\phi} \ , \qquad Z(\lambda,\phi) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\phi} \tag{3.68}$$

where $Z(\lambda,\phi)$ ensures normalization. For $\phi = 1$ it is the usual Poisson distribution, otherwise one has over-dispersion ($\sigma^2/\mu > 1$) for $\phi > 1$ or under-dispersion ($\sigma^2/\mu < 1$) for $\phi < 1$. The problem with the CMP distribution is that it is computationally very expensive to calculate the normalization term, which is why it was not feasible to use it for my calculations. However, it laid the basis for defining the following distribution whose dispersion can be varied by a parameter:

My distribution

From the experience with the CMP distribution, I searched for a distribution which does not involve an infinite summation in the normalization term. After some trial and error I chose

$$F(\lambda, k|\beta) = \mathrm{Gamma}(\lambda|k+1,\beta) = \frac{\beta^{k+1}}{\Gamma(k+1)} \lambda^k \, e^{-(\beta+1)\lambda}, \tag{3.69}$$

resulting in

$$P_{\mathrm{Kai}}(k|\lambda,\beta) := \frac{1}{Z(\lambda,\beta)} \frac{\beta^{k+1} \, \lambda^{2k} \, e^{-\beta\lambda}}{\Gamma(k+1) \, k!} \ . \tag{3.70}$$

The normalization term is according to Mathematica (Wolfram Research):

$$Z(\lambda,\beta) = \sqrt{\beta} e^{-(\beta+1)\lambda} \, I_1(2\sqrt{\beta}\lambda) \tag{3.71}$$

where $I_1(x)$ is the modified Bessel function of first kind [110]:

$$I_1(x) = \frac{1}{\pi} \int_0^\pi e^{x\cos\theta} \cos\theta \, \mathrm{d}\theta \ . \tag{3.72}$$

Another reason I chose that distribution is that unlike the Delaporte distribution but like the Poisson distribution there is a probability of one to have zero spikes if the rate is zero, $P(k=0|\lambda=0) = 1$, i.e. there are no spontaneous rates.

I will now use the Delaporte distribution and my distribution as output noise models and will compare them to the case of Poisson noise: for a given output noise strength λ and parameter(s) $(\alpha,)$ β the spike count of a neuron, k, will be distributed according to $P_{\mathrm{Poiss}}(k|\lambda)$, $P_{\mathrm{Delap}}(k|\lambda, \alpha, \beta)$, and $P_{\mathrm{Kai}}(k|\lambda, \beta)$. Then, their effects on maximum information and optimal thresholds will be compared.

Comparing the Delaporte and my distribution to the Poisson distribution

In Figure 3.26, one can see how the two modified Poisson distributions differ from the original Poisson for various parameter combinations of λ and β. For a small rate parameter λ (Fig. 3.26, left column) my distribution is shifted to $k = 0$ compared to the Poisson; in contrary the Delaporte is shifted to $k > 0$, however, this shift can be controlled with β. For a large λ (Fig. 3.26, right column) the Delaporte distribution is still always shifted to larger k compared to the Poisson, while my distribution is either shifted to smaller k (when β small) or larger k (when β large).

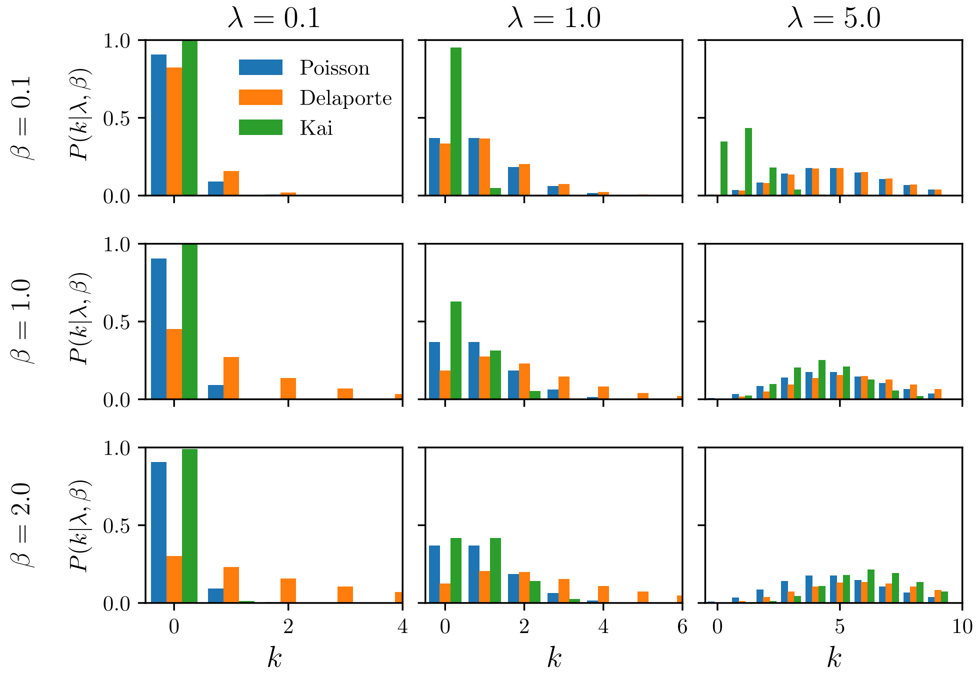


Figure 3.26: Comparison of the three discrete probability distributions $P_{\mathrm{Poiss}}(k|\lambda)$ (blue), $P_{\mathrm{Delap}}(k|\lambda, \alpha = 1, \beta)$ (orange), and $P_{\mathrm{Kai}}(k|\lambda, \beta)$ (green) for different rates λ along the horizontal direction and different parameters β along the vertical direction.

Figure 3.27 shows for various values of β how variance and mean depend on λ.

The Delaporte distribution shows over-dispersion for all values of $\alpha, \beta > 0$ which increases with increasing α (not shown) and β, whereas my distribution shows under-dispersion which increases with increasing β (deducted only from numerical calculations since I did not manage to calculate analytic expressions for the mean and the variance). For the Delaporte distribution the over-dispersion is present for all λ but technically decreases with increasing λ since both mean and variance depend linearly on λ. With my distribution, on the other hand, dispersion only becomes significant for $\lambda > 1$.

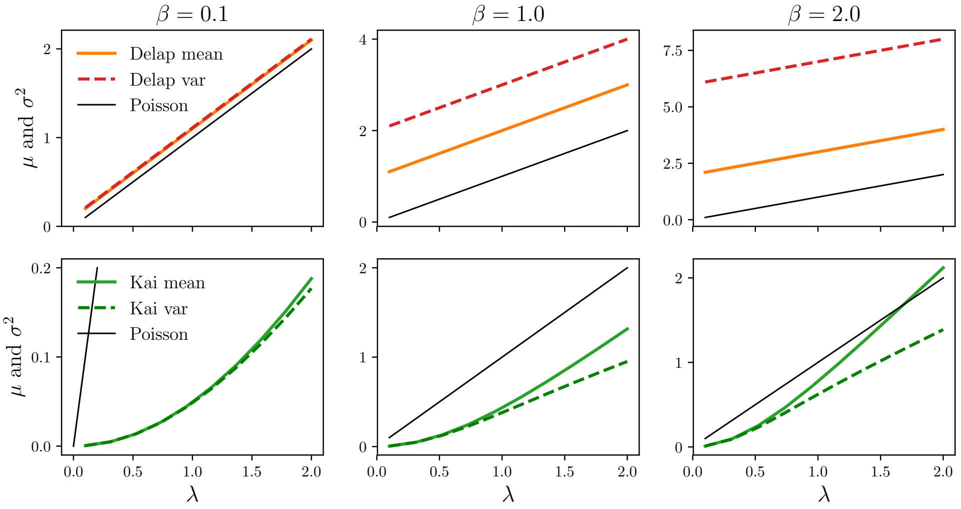


Figure 3.27: The Delaporte distribution is over-dispersed and has a spontaneous rate larger zero, while my distribution is under-dispersed and has a spontaneous rate equal to zero. Shown are the mean, μ, (solid) and variance, σ^2, (dashed) depending on rate parameter λ for Poisson and Delaporte distribution (top row) and my proposed distribution (bottom) for different values of additional parameter β. Values were obtained numerically. For the Delaporte distribution, $\sigma^2/\mu > 1$ and $\mu > 0$ for all $\beta > 0$ and any λ. For my distribution, $\sigma^2/\mu < 1$ for all $\beta > 0$ and, additionally, $\mu = 0$ for $\lambda = 0$. Note the different scaling of the y-axes. The α parameter of the Delaporte distribution was set to one.

For the three noise models (Poisson, Delaporte, Kai) I compared information encoding and optimal thresholds for a system of two binary neurons in dependence of the output noise level, R, where setting R as the rate parameter λ. As before, smaller R means larger output noise.

Comparing the effect of Delaporte noise to Poisson noise, the maximized information is decreased and optimal thresholds are shifted back towards the stimulus mean for given output noise parameter R (Fig. 3.28). For small β, the Delaporte indeed resembles the Poisson distribution as long as R is not to close to zero (Fig. 3.28A,B).

For significant β, however, the maximized information is reduced and thresholds are at the stimulus mean for $R \to 0$ (Fig. 3.28C-F). The information loss is in accordance with the fact that the effective output noise at any given R is larger for Delaporte noise compared to Poisson noise (Eqs. 3.66, 3.67; Fig. 3.27). Similarly, the shift of optimal thresholds towards the stimulus mean is in accordance with the fact that Delaporte noise causes spontaneous rates, since $P_{\text{Delap}}(k_i > 0|\nu_i = 0) > 0$, such effectively qualitatively reproducing the results of Section 3.6.4.1. However, the peculiar phenomenon that optimal thresholds are distinct for all output noise parameters $R > 0$ occurs also for Delaporte output noise.

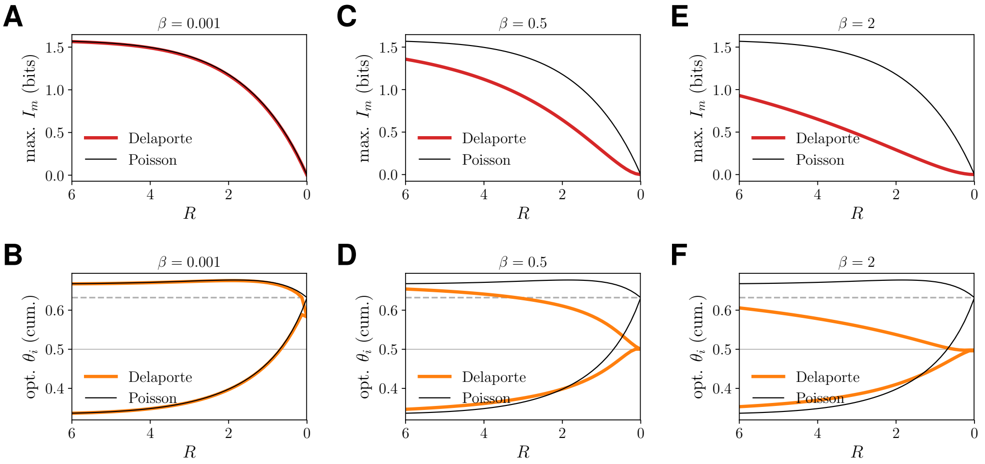


Figure 3.28: The Delaporte distribution as output noise model shifts optimal thresholds to the mean of the stimulus distribution. A. Maximized information depending on output noise R using the Delaporte distribution as output noise model with very small $\beta = 0.001$ ($\alpha = 1$ for all plots). For such small β, Delaporte noise seems to have the same effect on the information quantity as Poisson noise. **B.** Optimal thresholds in cumulative stimulus space leading to the information shown in A. The thresholds are same as with Poisson noise, except for slight deviations at large noise, $R \to 0$. The dashed horizontal line indicates the analytically calculated value of optimal thresholds for $R \to 0$ in the case of Poisson noise [106]. **C.** As A but with $\beta = 0.5$. For this parameter value, the Delaporte distribution causes more noise than standard Poisson noise, what leads to smaller information. **D.** As B but with $\beta = 0.5$. Compared to the Poisson, optimal thresholds are shifted to the stimulus mean of 0.5 (grey horizontal line) for $R \to 0$. **E,F.** As C,D but with $\beta = 2$. Information loss and threshold shifiting to the stimulus mean are even more pronounced.

Finally, I was interested in the case of having an over- or underdispers distribution as output noise model without spontaneous rate. Comparing the effect of my noise to Poisson noise, the maximized information is decreased for small β (Fig. 3.29A), but either decreased or increased for larger β (depending on output noise parameter R, Fig. 3.29C). For all β, the respective optimal thresholds still take the same value for $R \to 0$ as with Poisson output noise, namely $1 - 1/e$ (Fig. 3.29B,D).

Note, that even though my distribution shows under-dispersion for all $R > 0$, the information for a given R is not necessarily increased. The reason becomes obvious if one looks at the distribution (e.g. for $\lambda = 1$, $\beta = 0.1$ in Fig. 3.26): even for $R \gg 0$ the probability is high that zero spikes are emitted, thus leading to a high noise entropy and low mutual information. For larger values of R and β the effect of under-dispersion takes over (mostly because then zero spikes are very unlikely resulting in small noise entropy) and my distribution shows higher information compared to the Poisson. Thus, there is a nonlinear scaling of R compared to the Poisson distribution, since the noise is larger than with the Poisson for large R but smaller for small R.

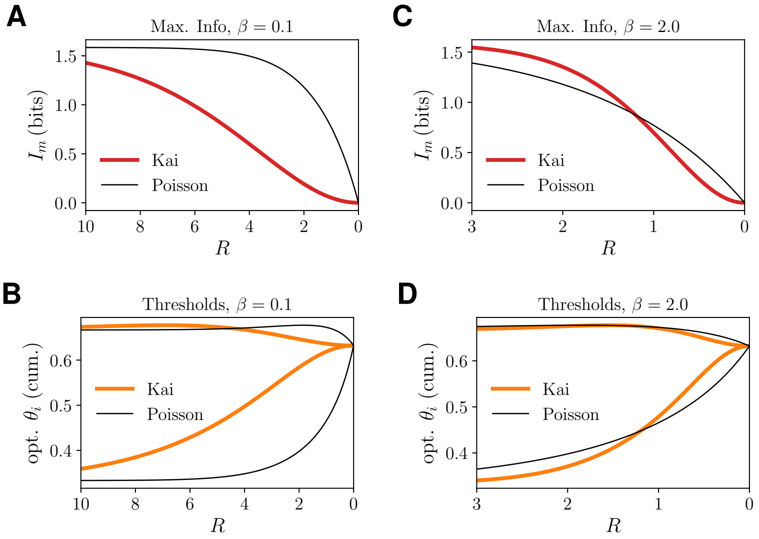


Figure 3.29: Without spontaneous rates there is no shift of optimal thresholds towards the stimulus mean. A. Maximized information depending on output noise R using my distribution as output noise model with $\beta = 0.1$ as parameter. For this parameter my distribution causes more noise then standard Poisson noise. **B.** Optimal thresholds in cumulative stimulus space corresponding to A. There is no shift towards the stimulus mean of 0.5 for $R \to 0$. **C,D.** As A,B but with $\beta = 2$. Note different scaling of x-axes.

Conclusion

The specific properties of the Poisson distribution of deterministic output at zero rate and normal dispersion (i.e. the mean being equal to the variance) can only partly explain the peculiar phenomena of the output noise: the phenomenon of optimal thresholds shifted away from the stimulus mean to $1 - 1/e \approx 0.63$ occurs in fact only in the case of deterministic output at $R \to 0$. However, the phenomenon that optimal thresholds are distinct for all $R > 0$ seems to occur irrespective of deterministic output at zero rate or dispersion, but instead, seems to be caused by stochastic spike generation, i.e. the nature of a discrete probability distribution.

3.7 Effect of the stimulus distribution

Remark: The method of this subsection has been mentioned in short in [1] and Figure 3.30 has been published in the supplementary material of [1]. Here, I describe the method and the results in a detailed manner.

Throughout this work, I investigate the encoding of a one-dimensional stimulus drawn from a Gaussian distribution; however, natural stimulus distributions have a higher level of sparseness than the Gaussian distribution [111, 112]. Therefore, I explored information maximization using the GND as the stimulus distribution. In analogy to the input noise distribution, using the GND allows me to continuously vary the kurtosis (Sec. 3.6.1.1). Otherwise, the setup is the same as in the beginning of the chapter: I use the independent-coding channel with a population of three neurons with Poisson output noise and Gaussian input noise. The noise levels are quantified by the expected spike count at maximum firing rate, $R = \nu_{\max}\Delta T$, and the standard deviation of the input noise distribution, σ. The standard deviation of the stimulus distribution is still set to one (remember that increasing the standard deviation of the stimulus distribution is equivalent to decreasing the standard deviation of the input noise distribution), while I varied the kurtosis by changing the parameter β of the GND (Eq. 3.42). Qualitatively, my results did not change much (Fig. 3.30) compared to using the Gaussian stimulus distribution ($\beta = 2$, Fig. 3.4A). For increased kurtosis of the stimulus distribution ($\beta = 1$, i.e. a Laplace distribution), there are still subsequent threshold bifurcations with respect to both noises, however, the range of intermediate threshold diversity with two thresholds being equal is quite small for small σ (beige color in Fig. 3.30A). Since there are no "bumps" visible anymore in Figure 3.30A the non-monotonic behavior has disappeared for such a large kurtosis of the stimulus distribution. On the other hand, for small kurtosis ($\beta = 7.5$, resembling more a uniform distribution), the non-monotonic behavior is still present, what is indicated by the bumps in Figure 3.4B. The conclusion is that the shape of the stimulus distribution does not play a significant role in my framework. Also the non-monotonic threshold behavior is not restricted to the stimulus distribution being Gaussian.

3.8 Summary and discussion

In this chapter, I maximized the mutual information between stimulus and responses of a population of neurons with binary nonlinearities. The stimulus is corrupted by two different noise sources, namely additive input noise before the nonlinearities and Poisson output noise after the nonlinearities. I compared two scenarios for stimulus processing commonly used in previous studies, specifically, encoding the stimulus with independent transmission channels and lumping the channels into one effective channel. In each scenario, I calculated the maximized mutual information and the optimal thresholds of the population and portrayed

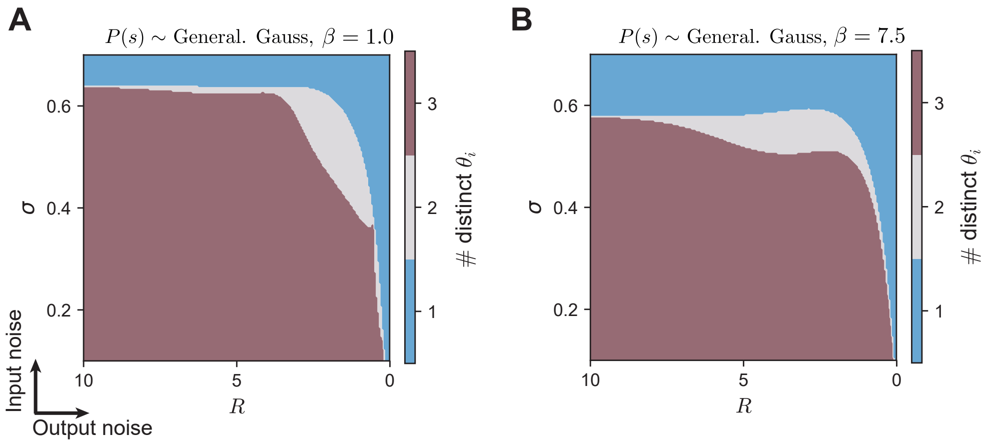


Figure 3.30: Number of distinct optimal thresholds when varying the kurtosis of the stimulus distribution through using the generalized normal distribution with varying β. Independent-coding channel with a population size of three neurons. **A.** Laplacian (having high kurtosis) as input distribution. **B.** Input distribution with low kurtosis (similar to uniform). From [1].

how they depend on the strength and shape of the two noise sources. For some noise parameters, I obtained unexpected non-monotonous behavior of the optimal thresholds.

Lumping channels causes information loss

Unsurprisingly, increasing either input or output noise in the population decreases the total amount of transmitted information. For all finite noise levels, the independent-coding channel always encodes more information than the lumped-coding channel, especially for biologically realistic, intermediate output noise values (Sec. 3.3, Fig. 3.2). This occurs because lumping multiple information pathways into a single coding channel reduces the possible values of the encoding variable and increases the noise entropy and thus introduces additional noise (Fig. 3.3). Therefore, threshold bifurcations in the lumped-coding channel occur at significantly lower critical noise levels compared to the independent-coding channel (Sec. 3.4, Fig. 3.4).

I only treated the extreme cases of full lumping – where the outputs of all neurons are lumped into a single variable – and no lumping. In principle, different combinations of partial lumping, e.g. lumping three outputs into two channels, are also possible. Partial lumping is a common strategy in sensory systems [113]. Furthermore, I assumed no weighting of inputs during the lumping process. This is an oversimplification since in biology spikes from different presynaptic neurons usually very differently impact the membrane potential of the postsynaptic neuron. These individual weights could also be optimized [86].

Optimal thresholds bifurcate successively from no to full redundancy

In general, the number of distinct optimal thresholds decreases with increasing noise of either kind at critical noise levels by successive bifurcations of the optimal thresholds (Sec. 3.4). This is the case for both channel types, however, for the independent-coding channel the bifurcations are continuous, while for the lumped-coding channel, the bifurcations are discontinuous (Fig. 3.4). Interestingly, in the case of the lumped-coding channel with only one noise source, threshold bifurcations become continuous. It remains unclear what the exact reasons for the difference in continuity are.

Another interesting point for populations of four or more neurons is the phenomenon of a "subtle" increase in redundancy, where the optimal number of distinct thresholds does not change but their allocation does (e.g. optimal thresholds changing from $\theta_1 = \theta_2 = \theta_3 \neq \theta_4$ to $\theta_1 = \theta_2 \neq \theta_3 = \theta_4$). This change happens in a discontinuous manner even for the independent-coding channel (Fig. 3.5).

Unexpected non-monotonous behavior of optimal thresholds

Interestingly, for a small range of noise parameters, I found a non-monotonic change in the number of distinct optimal thresholds with noise levels (Sec. 3.5, Fig. 3.12). After having made considerable efforts to verify and understand this behavior it seems indeed to be a true phenomenon, however, an intuitive explanation for this behavior is still missing. The only similar result that I have found in the literature is a non-monotonicity of optimal threshold diversity when maximizing Fisher information (a local information measure) for neurons encoding sound direction [84]. There, the optimal thresholds diversity is first increasing, then decreasing and finally increasing again with encoding precision, however, the authors did not comment on this non-monotonicity. When looking at optimal threshold differences instead of optimal number of distinct thresholds, I see a non-monotonic behavior for a very large range of noise parameters. The two phenomena are very probably related and understanding one could help in understanding the other.

The influences of the noise sources

For finite noise levels, both Gaussian input noise and Poisson output noise seem to have the same effect on threshold bifurcations, i.e. there is no qualitative difference between varying one noise source or the other. For the limit of one noise vanishing or being very strong, however, distinct effects become apparent regarding the symmetry of the optimal thresholds and regarding subsequent transitions from zero to full redundancy. Moreover, the exact shape of the input noise distribution does not matter, as least as long as it is symmetric (Sec. 3.6.1, Fig. 3.18). Thus, in future studies, it would be interesting to investigate if and how skewed input noise distributions qualitatively affect optimal thresholds. For the case of neurons being exposed to distinct noise levels, I confirmed for two neurons that the information landscape becomes asymmetric (Sec. 3.6.2). However, the direction of asymmetry is not as trivial as previously reported [29] as it depends on the noise levels: for the overall noise being small it is optimal that the neuron with *high* input noise has the lower threshold, while for the overall noise being high it is optimal that the neuron with *low* input noise has the smaller threshold (Fig. 3.22-3.24). For the output noise, it became apparent that the presence of spontaneous rates has a clear effect on the asymmetric shift of optimal thresholds towards larger values (Sec. 3.6.4). Unfortunately, due to the computational expensiveness of incorporating spontaneous rates, it was hard to study this effect on a population with more than two neurons.

4 Information landscapes and phase transitions at critical noise values

Remark: Some of the methods, results, figures, tables, and text in this chapter are part of an article entitled *Efficient population coding depends on stimulus convergence and source of noise* which has been written together with Shuai Shao and Julijana Gjorgjieva. The article has been uploaded to the preprint server *bioRxiv* [1] and is currently under review for publication in the journal *PLOS Computational Biology*. All methods, results, figures, tables, and text from that article which are part of this chapter were my contribution to the article. In this chapter, I specifically mention if and what I have taken from [1] before each (sub)section.

In the previous chapter, I have described extensively how the optimal thresholds and the maximized mutual information depend on noise levels. An interesting question is about how much information is lost when using suboptimal compared to optimal thresholds. In the current chapter, I, therefore, want to investigate how important it actually is that thresholds are precisely optimized, i.e. to quantify information loss when thresholds (slightly) deviate from their optimal values. If suboptimal thresholds cause negligible information loss then there is no strong incentive for biological systems to achieve near-optimal threshold values during evolution. Then, my predictions would unlikely be confirmed in experiments. Therefore, in this chapter, I will first look at the shape of the information landscape (Sec. 4.1). Since it is hard to visualize the shape of high-dimensional landscapes I will use methods from differential geometry. That way I find that the information landscape takes a characteristic shape at critical noise values (Sec. 4.3). Furthermore, I quantify how much information is encoded if thresholds are sampled randomly (Sec. 4.2). The intention is to find out how much worse the information encoding with randomly chosen thresholds is compared to optimal thresholds and to understand how much random mutations of thresholds affect information encoding. Then, I will describe how information encoding with an independent- or lumped-coding channel undergoes phase transitions at critical noise levels that are well-described in physics (Sec. 4.4). Interestingly, this phenomenon of phase transitions can be related to the previous sections of this chapter, i.e. the fact that the information landscape takes specific shapes around critical noise values.

4.1 Using principal curvatures to quantify the sensitivity of information landscape

Remark: One method of this section (using the Hessian matrix of the information landscape, Eq. 4.3) has been mentioned in [1]. Here, I derive and describe this method in detail and discuss it using new figures (Figs. 4.1 and 4.2) to increase the understanding and limits of the method.

For a single neuron, the information just depends in a concave manner on the threshold. Also for a population of two or more neurons, I found that the information is a (locally) concave function with respect to each threshold component. Thus, the idea could be to look at information loss ΔI_m when perturbing an optimal threshold component by $\Delta\theta_i$ (Fig. 4.1A). However, the functions $I_m(\theta_i)$ are in general different for each component i, and furthermore, I am also interested in quantifying how the information behaves when changing several components simultaneously. One idea is to identify those directions in threshold space along which the information landscape exhibits largest and smallest curvature (Fig. 4.1B). This would give upper and lower limits, respectively, of information loss due to threshold perturbations by a given magnitude. Another idea is to perturb the optimal threshold vector $\vec{\theta}^*$ by an arbitrary (but small) $\Delta\vec{\theta}$ and look at the information difference:

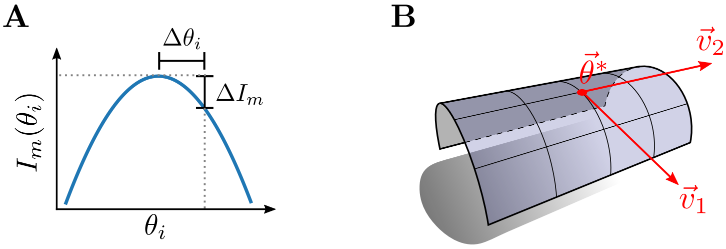


Figure 4.1: Quantifying information loss due to suboptimal thresholds using principal curvatures of the information landscape. A. Schematic showing how the mutual information I_m is a (locally) concave function with respect to each threshold component θ_i. Perturbing one component of the optimal threshold vector by $\Delta\theta_i$ causes an information loss ΔI_m. **B.** Schematic of the information landscape for $N = 2$ neurons for which the curvature is highly asymmetric: the information loss due to perturbing optimal threshold vector $\vec{\theta}^*$ can vary substantially between perturbations in the direction of largest or smallest curvature (denoted as $\vec{v}_1$ and $\vec{v}_2$, respectively). These directions are the eigenvectors of the Hessian matrix of the information and called “principal curvatures”. Adapted from [114].

$$\Delta I := I(\vec{\theta}^*) - I(\vec{\theta}^* + \Delta\vec{\theta}) \tag{4.1}$$

In the following, I will show that these two ideas are interconnected since they base on the concept of *principal curvatures*, which is a well-known concept in differential geometry [100]. Note that for simplicity from now on I drop the subscript of the mutual information and set $I \equiv I_m$. Expanding Eq. 4.1 around $\vec{\theta}^*$ up to second-order one obtains

$$I(\vec{\theta}) \;\approx\; I(\vec{\theta}^*) \;+\; \nabla I(\vec{\theta}^*)\,(\vec{\theta}-\vec{\theta}^*) \;+\; (\vec{\theta}-\vec{\theta}^*)^T\,\mathcal{H}I(\vec{\theta}^*)\,(\vec{\theta}-\vec{\theta}^*)\;. \tag{4.2}$$

The second term vanishes since the gradient at the maximum, $\nabla I(\vec{\theta}^*)$, is zero. The third term contains the Hessian matrix $\mathcal{H}$ of the information, which is defined as [115]

$$(\mathcal{H}I)_{ij} = \frac{\partial^2 I(\vec{\theta})}{\partial\theta_i \partial\theta_j}\;. \tag{4.3}$$

$(\mathcal{H}I)_{ij}$ is a symmetrical matrix and performing an eigendecomposition yields the directions of *principal curvatures* as eigenvectors and the respective magnitude of the curvatures in these directions as eigenvalues [100].[1] These principal curvatures provide a lot of information on the curvature of the information landscape. In two dimensions, for example, the principal curvatures at a point of a surface are the directions of largest and smallest curvature (see again Fig. 4.1B which shows the two principal curvatures of a cylindrical object). For my general case of having an N-dimensional information landscape, $I(\theta_1, ...\theta_N)$, the N principal curvatures give the directions in threshold space along which the curvature of the information landscape is of special interest. Note that from a mathematical perspective, the information landscape is an N-dimensional manifold in $N+1$-dimensional space [100]. Taking together Eqs. 4.1 and 4.2 one obtains for the information loss:

$$\Delta I \approx -(\Delta\vec{\theta})^T\;\mathcal{H}I\;(\Delta\vec{\theta}). \tag{4.4}$$

My first attempt to systematically quantify the information loss due to suboptimal thresholds was to compare the relative information loss, $\Delta I/I_{\max}$, when perturbing the optimal threshold vector $\vec{\theta}^*$ by magnitude δ in a direction of principal curvature:

$$\frac{\Delta I}{I_{\max}} = \frac{I(\vec{\theta}^* + \delta\vec{v}_1)}{I_{\max}}\;. \tag{4.5}$$

This is schematically illustrated in Figure 4.2A. From my experience with spotting and avoiding local maxima (Sec. 3.4.4), I suspected that the relative information loss scales with the population size and indeed there was some dependency of the relative information loss on the neuron number N (Fig. 4.2B-D). For most noise level combinations, the relative information loss is highest for $N=1$ and decreases with increasing N (Fig. 4.2D). However, the relative information loss strongly depends on the noise levels (compare Fig. 4.2B with C). Moreover, one can see that this

[1] In differential geometry a principal curvature of a surface is defined as the direction along which the normal vector (being perpendicular to the surface) only moves in a plane spanned by the principal curvature and the normal vector if one moves into the direction of this principal curvature.

approach is too simple in many cases: for $N = 2$ and two different values of output noise R, plotting the information landscape shows that the direction of greatest curvature points to the other maximum[2] (Fig. 4.2E,F). Especially, in this case, there might be a big difference between the positive and negative direction of $\vec{v}_1$ (though this difference usually depends on the strength of the perturbation, δ). Additionally, in this case, the information loss is strongly non-monotonic with δ. Thus, the approach of Eq. 4.5 only makes sense locally, i.e. within a very small distance from the maximum. Furthermore, it is difficult to compare perturbations across different population sizes: If, for example, the total perturbation is of length δ along the diagonal of the threshold space, then the perturbation in each dimension is $\delta/\sqrt{N}$, i.e. each threshold component is perturbed to a smaller extend the higher the dimension N.

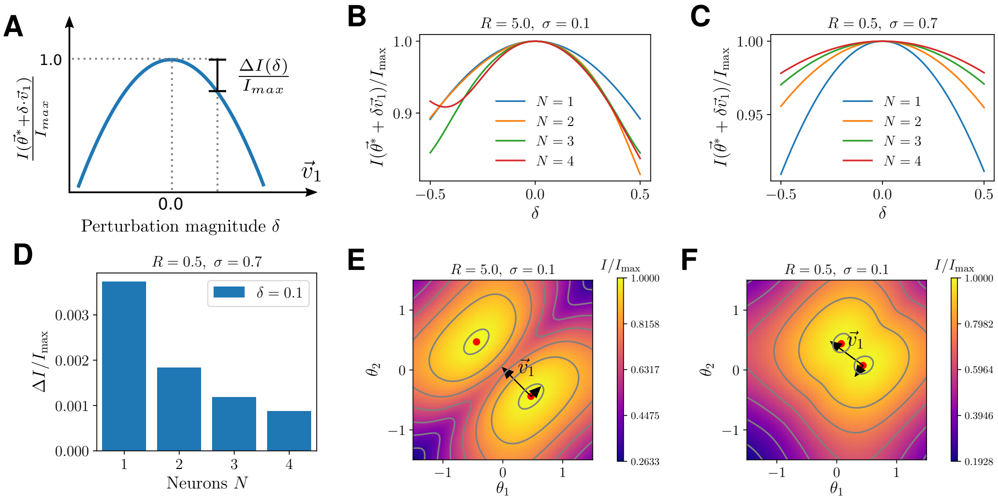


Figure 4.2: The attempt to quantify information loss along the largest curvature of the information landscape. A. Schematic showing qualitatively how the information normalized by the maximum information decreases when perturbing optimal thresholds by magnitude δ in the direction of largest curvature of the information landscape $\vec{v}_1$. **B.** Actual information when perturbing optimal thresholds for $N = \{1, 2, 3, 4\}$ neurons, each normalized by maximum information for the respective N. Independent coding-channel with the case of low noise levels ($R = 5$, $\sigma = 0.1$). **C.** As B but with high noise levels ($R = 0.5$, $\sigma = 0.7$). **D.** Normalized information loss for each N when perturbing optimal thresholds by going $\delta = 0.1$ threshold units away from the optimum in the direction of largest curvature. **E, F.** Normalized information landscapes in the case of $N = 2$ neurons for the noise values used in B and C, respectively. The direction of largest curvature could point towards to or away from the other maximum, which could be further or closer away.

The conclusion is that the Hessian matrix of the information landscape and its eigenvectors and eigenvalues quantify the local curvature of the information land-

[2] The two maxima are equivalent due to symmetry reasons: having two neurons with thresholds θ_1, θ_2 is equivalent to having two neurons with thresholds θ_2, θ_1 as long as both neurons are exposed to the same noise levels (see Sec. 3.6.2). Thus, both maxima are "the" global maximum.

scape. However, since this is a local measure, it can only be used to quantify information loss for suboptimal thresholds in very close proximity to the optimal thresholds. Thus, these information losses are naturally small and cannot be used to gain systematic insights on how much information is lost when suboptimal thresholds are more distant from the optimal thresholds. In Section 4.3, however, I demonstrate how the eigendecomposition of the Hessian matrix still provides very valuable information about the shape of the information landscape. First, I will describe a more suitable approach of systematically quantifying information loss due to suboptimal thresholds in the next section.

4.2 Information loss for randomly sampled thresholds

In the previous section, it became apparent that when perturbing the optimal thresholds the information loss seems to decrease as the number of neurons in the population increases. Indeed, increasing the number of neurons in the population reduces the contribution of each individual threshold to the total information. To go beyond the local analysis performed in the previous section, I compared the information loss (relative to the optimal information I_{max}) achieved in populations where the thresholds of the neurons were chosen randomly. It became apparent, that there are at least three different ways of randomly choosing ("sampling") thresholds: First, by sampling each threshold component from the stimulus distribution, second, by sampling each component around the optimal threshold vector, and third, by sampling around the optimal threshold vector but correcting the distance between the sampled threshold vector to the optimal threshold vector by taking into account the dimensionality of the threshold vector.

Each of these three approaches has its supporting arguments for why each is the right one: with the first approach only the stimulus distribution is taken into account but no information about the information landscape with its optimalities. It would resemble the "beginning" of an evolutionary process. The second approach of sampling around the optimum gives insights into how much information is lost when perturbing optimal thresholds, i.e. the evolutionary pressure to stay at or very close to the optimum, for example in the presence of mutations. The third approach has the same rationale but explicitly takes into account that the distance between the optimal threshold and the sampled threshold in the second approach increases with population size N. Thus, the sampled thresholds in the third approach are corrected for the dimension such that on average the distance between optimal and sampled threshold are equal for all population sizes. In the following, I describe the implementation and the results for each approach. All of this will be done for the independent-coding channel for computational tractability and to be able to go for N as large as four.

4.2.1 Sampling thresholds around null-vector

The first approach is to sample each threshold component independently from the stimulus distribution $\mathcal{N}(0,1)$, which means that the whole threshold vector is sampled from a multi-variate Gaussian with unit variance:

$$\vec{\theta} \sim \mathcal{N}(\vec{0}, \mathbb{1}) \; . \tag{4.6}$$

Figure 4.3A,B shows the relative information loss as distributions for $N \in \{1,2,3,4\}$ for high and low noise levels. While it is common for $N = 1$ to have a very small relative information loss close to zero and also not uncommon to have very high relative losses close to 1, both become increasingly uncommon with increasing population size N. This is expected, since for very small or very large information losses, each component of the optimal threshold has to be close to its optimal value or very far away from the stimulus mean, respectively. With increasing N, however, it becomes much less probable to sample threshold vectors where *each* component is very close to the optimal value or very far away from the stimulus mean. With increasing population size the relative information loss is slightly decreasing (29% for $N = 1$ vs. 24% for $N = 4$, Fig. 4.3G). This qualitative result was independent of the exact values of the two noise sources (not shown).

4.2.2 Sampling thresholds around optimal threshold vector

Second, I sampled each threshold component around the optimal threshold vector to quantify how much information is lost when the optimal thresholds are perturbed. Again, I choose a multi-variate normal distribution with unit variance, but this time the mean is the optimal threshold vector:

$$\vec{\theta} \sim \mathcal{N}(\vec{\theta}^{*}, \mathbb{1}) \; . \tag{4.7}$$

This could help understand the size of the evolutionary pressure to maintain optimality in the presence of genetic drift. Qualitatively, both the distributions (Fig. 4.3C,D) and the mean (Fig. 4.3H) of the relative information loss show the same behavior as with the previous sampling method. The relative information loss slightly decreases with N (24% for $N = 1$ vs. 20% for $N = 4$).

For both of these two ways of sampling thresholds, i.e. sampling around the origin and around the optimum, the average Euclidean distance $\langle d \rangle$ between the sampled threshold vectors and the optimal threshold vector $\vec{\theta}^{*}$ increases with the dimension N of the threshold vector:

$$\langle d \rangle = \left\langle \sqrt{\sum_{i=1}^{N} (\theta_i - \theta_i^{*})^2} \right\rangle_{i=\{1,..,N\}} \quad \sim \quad \sqrt{N} \; . \tag{4.8}$$

Even more impactful, the distribution of the distance becomes extremely skewed away from small distances with larger N. This phenomenon is often called *the curse*

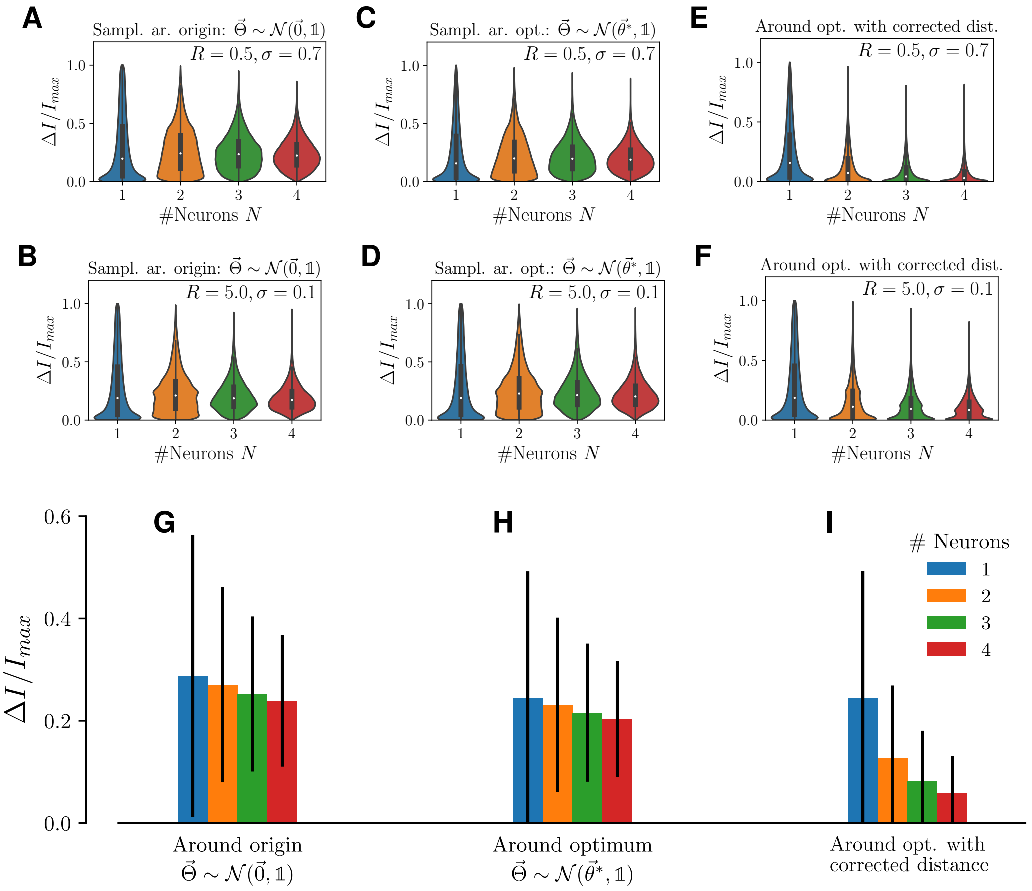


Figure 4.3: Information losses for randomly sampled thresholds compared to optimal thresholds decreases with neural population size. For the independent-coding channel and normalized by maximum information for each neuron number N. **A,B.** Sampling thresholds around the origin (the null-vector) by drawing each component of the threshold vector from a Gaussian distribution $\mathcal{N}(0, 1)$ (same as the stimulus distribution). This is done for $N = \{1, 2, 3, 4\}$ neurons (independent-coding channel only) and then the resulting normalized information distributions are shown for high noise levels (A) and low noise levels (B). **C, D.** As in A,B but instead of sampling around the origin, the threshold vectors for each N are Gaussian sampled around the respective optimal threshold vector $\vec{\theta}^*$: $\vec{\Theta} \sim \mathcal{N}(\vec{\theta}^*, \mathbb{1})$. **E, F.** As in C,D but in such a way that the sampled threshold vectors are corrected by length, i.e. have on average the same distance to the optimum for each N (see text). **G-I.** Mean and standard deviation of the relative information losses shown in A,C, and E, respectively, i.e. each for high noise.

of dimensionality [116]. For example, the probability to be in the range of $[-0.1, 0.1]$ when sampling one threshold from $\mathcal{N}(0, 1)$ is approximately 0.08. When sampling N thresholds from $\mathcal{N}(0, 1)$, the probability that *all* thresholds are in $[-0.1, 0.1]$ is only 0.08^N. If only *one* of the N thresholds is outside of $[-0.1, 0.1]$, then the whole threshold vector has already a length > 0.1. Thus, it is not so unlikely to have a length of smaller 0.1 with $N = 1$, but extremely unlikely with $N = 4$ (namely $\approx 4 \cdot 10^{-5}$).

4.2.3 Sampling thresholds with corrected distance

Thus, in the third way of sampling thresholds I corrected for the length of the sampled values so that for all N the average distance of perturbation in threshold space is equal. This cannot be done by naively adapting the variance matrix, for example by sampling from $\mathcal{N}(\vec{\theta}^{*}, \mathbb{1}/N)$. Instead, one has to first, uniformly sample the direction from threshold space [117], and second, sample the distance to the optimal threshold vector from $\mathcal{N}(0, 1)$. This threshold vector is then added to the optimal threshold vector and the relative information loss is compared. I found in fact that the relative information loss decreases much stronger with N (24% for $N = 1$ vs. 6% for $N = 4$, Fig. 4.3I).

There is no obvious answer to which of the three measures[3] is the "right" one. On the one hand, the first two measures underestimate the decrease of relative information loss with N since the fact that the distance between sampled and optimal thresholds increases with N is not taken into account. On the other hand, one can argue that the third measure overestimates the relative information loss with N since in this case, the perturbation of each threshold component becomes smaller with N. Depending on the question one wants to answer, the importance of thresholds being precisely at the optimum decreases with N to different extents.

4.3 Characteristic shape of the information landscape at critical noise levels

Remark: The methods, results, figures (Figs. 4.4 and 4.5A), and text of this section have been originally published in [1] and content-wise are mostly identical to [1]. The differences are that I in this section I also describe results and show figures for larger population sizes (Fig. 4.5B,C) and non-monotonic thresholds (Fig. 4.6) and put it in context to the results from [1]. Apart from that, I have slightly modified some of the original text for stylistic reasons and increased clarity.

To gain a better understanding of the information landscape, especially at the critical noise values at which threshold bifurcations appear, I examined the Hessian matrix of the mutual information with respect to the thresholds (Eq. 4.3). To gain intuition about the differences of the information landscape between the

[3] The fourth possibility of sampling thresholds, namely with corrected length but around the origin, is not practicable since with larger N most threshold components will be sampled in very close proximity to the origin (so that the average distance to the origin is still one even with a larger amount of components). The information loss will then be highly dominated by the information loss that occurs when choosing the null-vector; in other words, the information loss will be dominated by $I(\vec{\theta}^{*}) - I(\vec{0})$ for larger N.

independent- and lumped-coding channels, I considered a population of two cells for which the landscape can be easily portrayed. However, I also showed that the theory extends naturally to populations with more neurons, i.e. information landscapes in higher dimensions.

I first considered the independent-coding channel for a fixed level of input noise, while varying the output noise. At the critical noise level, R_{crit}, where the thresholds bifurcate, one eigenvalue of the Hessian goes to zero (Fig. 4.4A). The information landscape undergoes a transformation around the critical noise levels, from a landscape with two distinct maxima separated by a local minimum at low noise, $R > R_{\mathrm{crit}}$ (Fig. 4.4B, top), where the population thresholds are distinct, to a landscape where there is a unique maximum at high noise, $R < R_{\mathrm{crit}}$, where the population thresholds are identical (Fig. 4.4B, bottom). For $R > R_{\mathrm{crit}}$, there are two inflection points (Fig. 4.4C, top), resulting in two opposite curvatures along the line that connects the two maxima. At the critical noise, $R = R_{crit}$, the two maxima converge at the bifurcation point and the two inflection points fuse together such that the curvature becomes zero (Fig. 4.4C, middle). At this point of convergence, the information landscape locally resembles a ridge, which extends along one principal direction of curvature (Fig. 4.4B, middle; see also Fig. 4.1B for the schematic). The ridge is perpendicular to the other principal direction, which stands for the direction of largest curvature. Finally, for $R < R_{\mathrm{crit}}$, the information landscape has a single maximum with negative curvature (Fig. 4.4B and C, bottom).

I then examined the eigenvalues of the Hessian matrix for a larger population of size $N > 2$. I found that at each critical noise level where the thresholds bifurcate, at least one eigenvalue of the Hessian matrix approaches zero. The number of zero eigenvalues – denoting the number of dimensions along which the information does not change locally – is equal to the number of thresholds participating in a bifurcation minus one. For $N = 3$, for example, there are two critical noise values at which the thresholds bifurcate (Fig. 4.5A,B, top). Therefore, at critical noise values where three thresholds are involved in the bifurcation the number of eigenvalues approaching zero is two, while at the critical noise values with only two thresholds involved thus the number of eigenvalues approaching zero is one (Fig. 4.5A,B, bottom). Similarly, for $N = 4$, three eigenvalues go to zero at the bifurcation where all four thresholds participate (Fig. 4.5C). Interestingly, for the bifurcation where only θ_3 and θ_4 participate, a different eigenvalue goes to zero compared to the bifurcation where only θ_1 and θ_2 participate.

Then I also looked at how the eigenvalues of the Hessian behaved for the nonmonotonic threshold behavior (Fig. 4.6). The same pattern as before emerged: at each bifurcation, the number of eigenvalues going to zero is the number of thresholds participating minus one. That is true for threshold ribbons (Fig. 4.6A,C), threshold switching (Fig. 4.6B) and also for threshold splitting with $N = 6$ (data not shown).

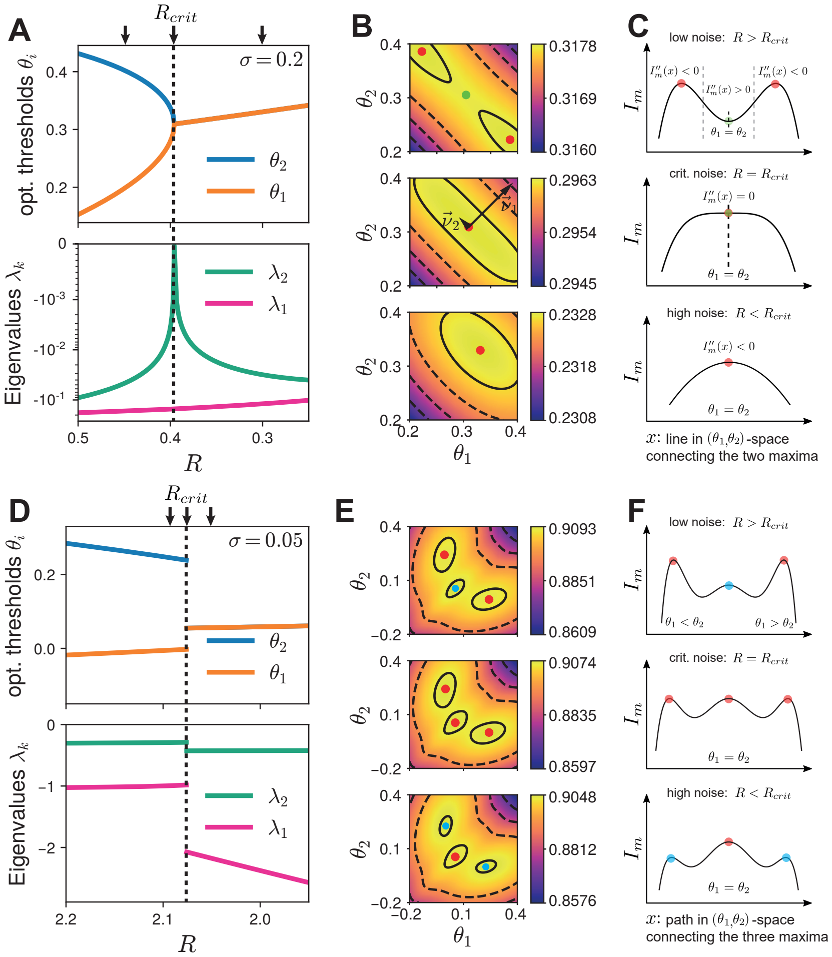


Figure 4.4: Information landscape for the independent- and lumped-coding channels behaves differently around critical noise levels. **A**. Top: Optimal thresholds of the independent-coding channel for a population of two neurons as a function of output noise R. Bottom: Corresponding eigenvalues of the Hessian of the information landscape with respect to thresholds. At the critical noise value $R_{crit} \approx 0.396$ at which the threshold bifurcation occurs (vertical dashed line) the smaller eigenvalue approaches zero. **B**. Information landscape $I_m(\theta_1, \theta_2)$ for the three output noise levels R indicated by arrows in A. Top: For $R > R_{\text{crit}}$, there are two equal maxima. Middle: At $R = R_{\text{crit}}$, the eigenvectors of the Hessian are shown and scaled by the corresponding eigenvalue (the eigenvector with the smaller eigenvalue, $\vec{v}_2$, was artificially lengthened to show its direction). At the critical noise value the information landscape locally takes the form of a ridge. Bottom: For $R \leq R_{\text{crit}}$, there is one maximum, meaning that the optimal thresholds are equal. **C**. The mutual information as a function of the line x in (θ_1, θ_2) space connecting the two maxima in B. Top: For $R > R_{\text{crit}}$ (low noise), there are two inflection points (dashed vertical lines) with zero curvature along the line x. The point with equal thresholds corresponds to a local minimum. Middle: At $R = R_{\text{crit}}$, the two maxima, the minimum, and the two inflection points merge in one point, thus the curvature is zero. Bottom: For $R \leq R_{\text{crit}}$, there is a single maximum with negative curvature. **D**. As in A but for the lumped-coding channel. Both the optimal thresholds and the eigenvalues show a discontinuity at critical noise level. The eigenvalues do not approach zero. **E**. Information landscape as in B for lumped-coding channel and noise values indicated by arrows in D. Local maxima are shown in cyan, global ones in red. **F**. Similar to C for lumped-coding channel. Here the abscissa denotes the (non-straight) path connecting the three maxima in E. From [1].

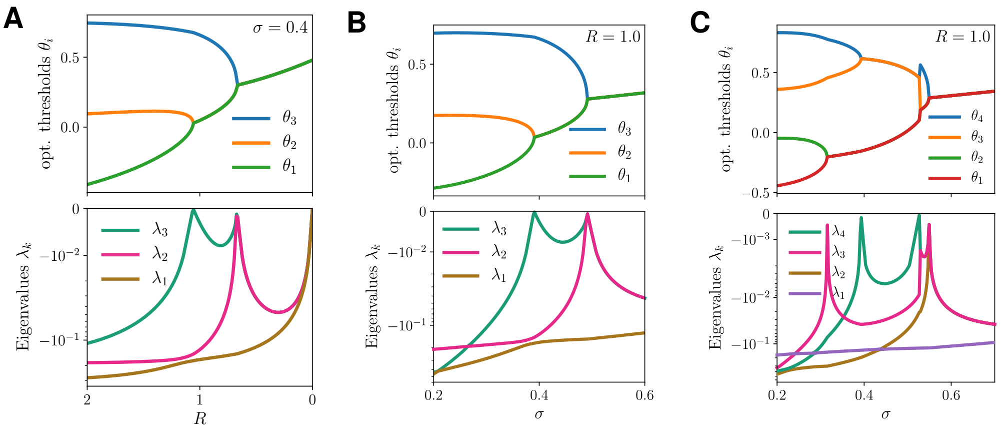


Figure 4.5: The number of eigenvalues of the Hessian matrix approaching zero is the number of thresholds participating in a bifurcation minus one. This is true with respect to both input and output noise. **A.** Optimal thresholds and corresponding eigenvalues depending on output noise R for $N = 3$ neurons and fixed input noise. **B.** As A but depending on input noise and with fixed output noise. **C.** As B but for $N = 4$ neurons. Panels A and B are from [1].

These results further support my claim that the aforementioned non-monotonic threshold behaviors are not an artifact.

Locally, threshold combinations along the ridge of the information landscape achieve almost the same information. Mathematically speaking, this ridge is a manifold [100] of dimension $M - 1$, where M is the number of thresholds involved in the bifurcation. Since the ridge is oriented at exactly 45° with respect to all of the θ-directions participating in the bifurcation the manifold is locally given by

$$\sum_{\left\{\substack{i \,\mid\, \theta_i \text{ involved} \\ \text{in the bifurcation}}\right\}} \theta_i = \text{constant}. \tag{4.9}$$

For example, for $M = 2$ this manifold is a line, while for $M = 3$ it is a plane and for $M = 4$ it cannot be visualized anymore. Following the same argument as for the population with $N = 2$ neurons (Fig. 4.4C,D), it can be shown that the curvature of the information landscape has to be zero in $M - 1$ principal directions, thus $M - 1$ eigenvalues of the Hessian have to be zero when M thresholds participate in a continuous bifurcation.

For the lumped-coding channel, the eigenvalues of the Hessian do not approach zero at the critical noise levels where the thresholds split (Fig. 4.4D). This is in agreement with the fact that threshold bifurcations are in general discontinuous for the lumped-coding channel (see also Fig. 3.4G,H). An exception to this is the limiting case when one noise level is zero, where the lumped-coding channel shows continuous bifurcations (Fig. 3.4I,J). At low output noise, $R > R_{\text{crit}}$ (Fig. 4.4E,F,

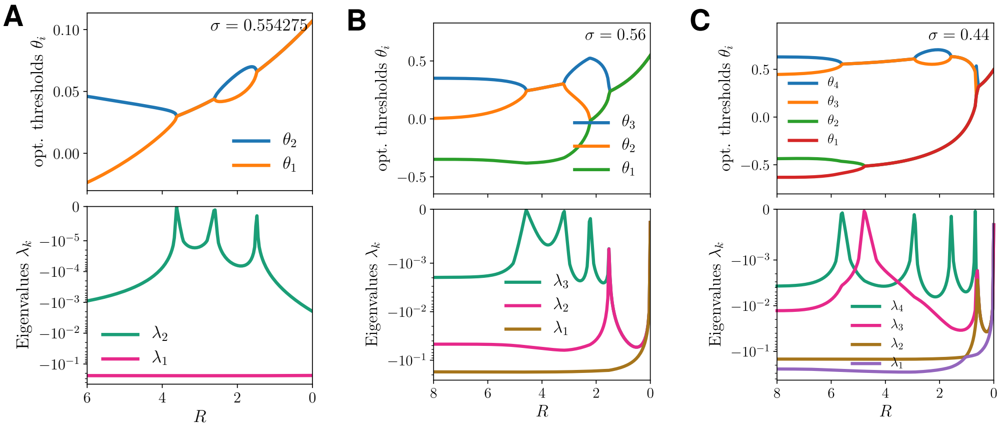


Figure 4.6: The same structure of eigenvalues of the Hessian matrix approaching zero at critical noise levels also holds true in the case of non-monotonic behavior of optimal number of distinct thresholds. **A.** Optimal thresholds and eigenvalues depending on output noise R for $N = 2$ neurons at a fixed input noise value σ for which the threshold ribbon appears. **B.** As A but for $N = 3$ with a threshold switching. **C.** As A but for $N = 4$.

top), the information landscape has two distinct maxima corresponding to the optimal thresholds, θ_1 and θ_2. However, the information landscape also has a local maximum at $\theta_1 = \theta_2$. As noise increases, this local maximum decreases more slowly compared to the two global maxima, until at the critical noise level R_{crit} the three maxima become equal (Fig. 4.4E,F, middle). As noise increases further, $R < R_{\text{crit}}$, the maximum at $\theta_1 = \theta_2$ becomes the single global maximum (Fig. 4.4E,F, bottom). In other words, the local maximum “overtakes” the global maximum at the critical noise value.

My results show, that for finite noise, the shape of the information landscape for the independent- and the lumped-coding channels can be uniquely related to the nature of the threshold bifurcations (continuous for the independent-coding and discontinuous for the lumped-coding channel). The information landscape takes a qualitatively different shape at the threshold bifurcations in each case, demonstrating the emergence of a new threshold through splitting either through a gradual “breaking” of the information ridge (Fig. 4.4B,C), or through a discrete switching from a local information maximum to the global maximum (Fig. 4.4F,E).

4.4 Phase transitions at critical noise levels

Remark: The methods and results (Fig. 4.7, 4.8, and 4.9 and the text) of this section (until including Subsec. 4.4.1) have been originally published in [1] and content-wise are identical to [1]. The differences are that in this section I have split up one figure into two (Fig. 4.7 and 4.8) for page width reasons. Furthermore, the methods, results, figures, tables, and text of Subsections 4.4.2 and 4.4.3 are taken from a modified version of [1] which currently is under review (see remarks at the beginning of Secs. 4.4.2 and 4.4.3). Apart from that, I have slightly modified some of the original text for stylistic reasons and increased clarity.

The characteristic bifurcations of the optimal thresholds at critical noise levels suggest the occurrence of phase transitions akin to those encountered in a variety of physical systems. In physics, a phase transition is defined by non-analytic behavior of the free energy – usually a discontinuity of its first or second derivative – and can be characterized by an order parameter [89]. For example, a phase transition occurs when the order parameter – which could among others be the magnetization of a ferromagnetic material – changes abruptly from zero to non-zero values with an external parameter, such as pressure or temperature. The quantities that change abruptly from zero to non-zero values in my diagrams of optimal thresholds are the threshold differences, which I thus treated as order parameters from now on. Furthermore, the external parameters are the two noise levels and the size that is potentially non-analytic is the mutual information. Guided by this characterization, I sought to relate the qualitative differences in optimal thresholds of the independent- vs. lumped-coding channel with two noise sources to phase transition phenomena. In specific, I wanted to test if there is indeed a close analogy between the following quantities: First, the mutual information being analogous to the free energy, second, the two noise sources being analogous to the temperature or pressure, and third, the threshold differences being analogous to the order parameter (as for example magnetization in a ferromagnetic material).

4.4.1 Bifurcations of optimal thresholds represent phase transitions

I illustrate the results for a population with three neurons. Using the analogy from physics, I have two order parameters which are the two threshold differences, $\theta_2 - \theta_1$ and $\theta_3 - \theta_2$. To determine whether a phase transition occurs, I computed the first and second derivatives of the mutual information with respect to a given noise parameter (Fig. 4.7). Using the classic Ehrenfest classification of phase transitions [91], a discontinuity in the first (second) derivative with respect to the noise implies a first- (second-) order phase transition. My approach also works with using the modern classification of phase transitions, which generalizes the second-order phase transition to any non-analytic behavior (i.e. discontinuous or divergent) of the information with a continuous first derivative [90]. There, the second-order phase transition is called a “continuous” phase transition since the order parameter is continuous as long as the first derivative of the information is continuous.

I found that the orders of the phase transitions always correspond to the discontinuity of the threshold differences – being the order parameters – when noise varied. For continuous threshold bifurcations, there was a discontinuity in the second derivative with respect to output noise while the first derivative was continuous,

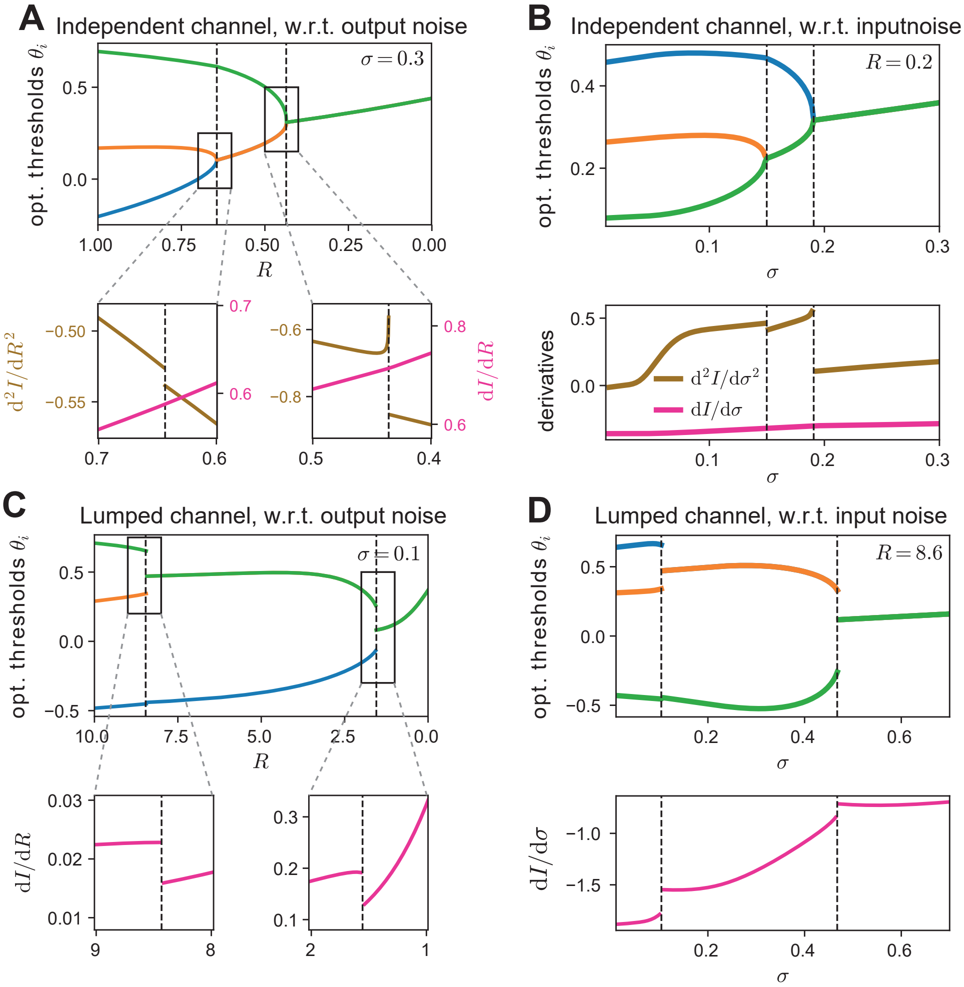


Figure 4.7: Bifurcations of optimal thresholds correspond to phase transitions with respect to both noise sources. To show the order of the phase transition the optimal thresholds and the derivatives of the mutual information with respect to both input and output noise are shown (both channel types, $N = 3$ neurons). **A**. Optimal thresholds for the independent-coding channel with respect to output noise R. Insets: The first derivative of the mutual information as a function of noise is continuous while the second derivative is discontinuous at the critical noise values where the thresholds bifurcate, implying a second-order phase transition. **B**. As in A, but with respect to input noise σ. **C**. Optimal thresholds as in A but for the lumped-coding channel. The first derivative is discontinuous at the critical noise values where the thresholds bifurcate, implying a first-order phase transition. **D**. As in C but with respect to input noise σ. Adapted from [1].

thus corresponding to a second-order phase transition (Fig. 4.7A). All phase transitions for the independent-coding channel were continuous and thus of second-order, i.e. also with respect to input noise (Fig. 4.7B). This result is in agreement with a previous study which also found a second-order phase transition in a population of two neurons in the presence of only input noise [29]; I extended this result to populations of more than two neurons and with more than one noise source. Next, I investigated phase transitions in the lumped-coding channel.

For discontinuous threshold bifurcations, I observed a discontinuity in the first derivative with respect to output noise and thus a phase transition of first-order (Fig. 4.7C). For the lumped-coding channel, almost all threshold bifurcations are discontinuous, i.e. first-order phase transitions; also when fixing output noise and varying input noise (Fig. 4.7D). The reason for that is that the bifurcation / phase transition happens at the noise level where the local maximum becomes the global one. This is a first-order phase transition since at this critical noise level the decrease of maximum information with noise changes abruptly, resulting in a discontinuity in the first derivative (Fig. 4.8H). An exception to this is when one noise source vanishes, e.g. when there is only input noise (Fig. 4.9B) or only output noise (Fig. 4.9D). There, most threshold bifurcations become continuous and thus also the phase transitions are of second-order. The single discontinuous bifurcation in Figure 4.9B stays first-order, as expected.

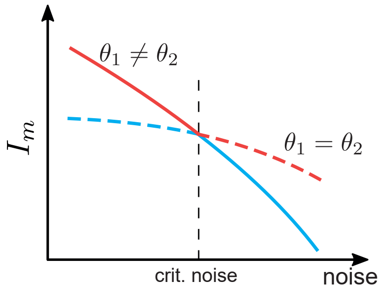


Figure 4.8: Schematic of why the discontinuous threshold bifurcations resemble first order phase transitions. At a discontinuous threshold bifurcation, the global maximum at $\theta_1 \neq \theta_2$ at low noise (red, solid) becomes a local maximum for high noise (cyan, solid), while $\theta_1 = \theta_2$ (dashed) becomes global. As their respective derivatives are different, there is a discontinuity in the first derivative when only taking the global maximum into account (red lines), corresponding to a first-order phase transition. From [1].

These results show that the threshold differences in the population of neurons resemble order parameters and determine the order of the observed phase transitions: discontinuous threshold differences correspond to first-order phase transitions while continuous threshold differences correspond to second-order phase transitions. In the following sections, I will investigate more similarities to statistical mechanics models from physics and also point out where my system shows peculiarities when compared to physics' systems.

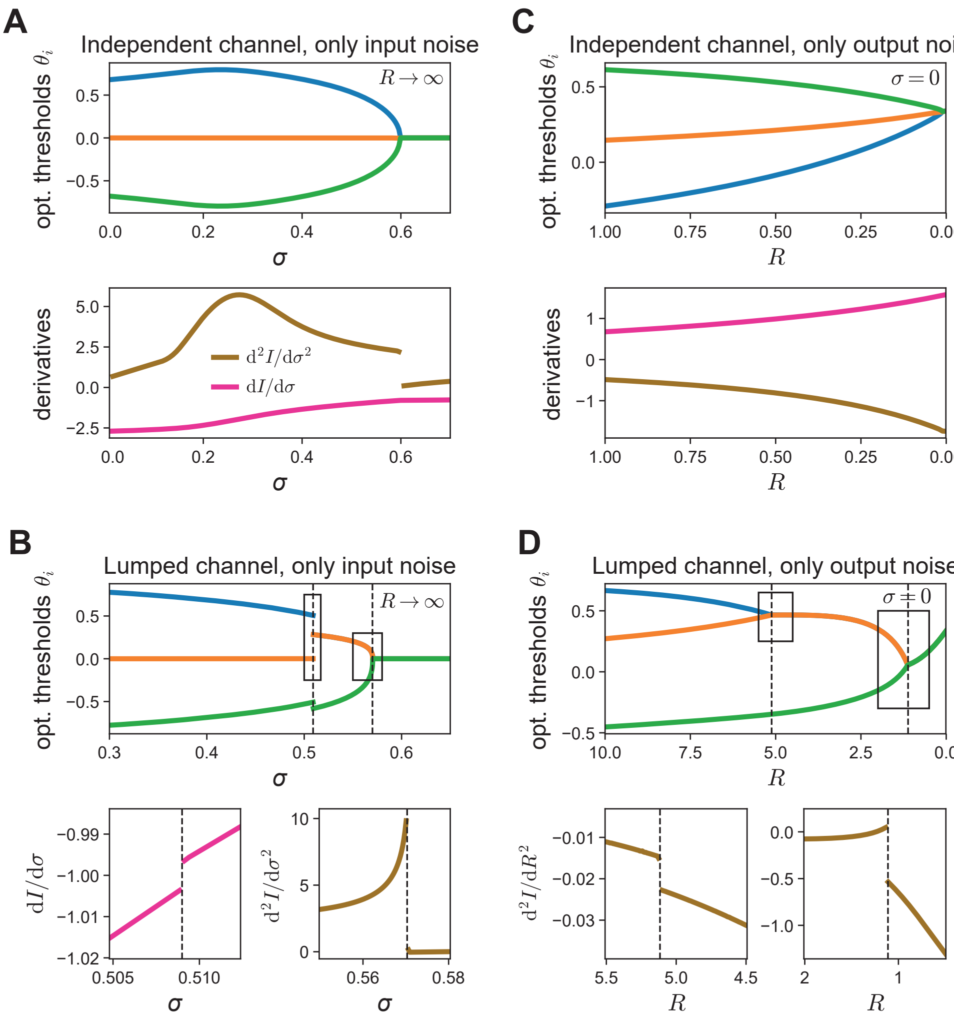


Figure 4.9: Bifurcations of optimal thresholds correspond to phase transitions in the limit of just one noise source. As in the previous figure, but now with $R = \infty$ or $\sigma = 0$, respectively. **A**. Threshold bifurcations for the independent-coding channel with respect to input noise σ for vanishing output noise. The derivatives of mutual information with respect to input noise indicate a second-order phase transition. **B**. As in A but for lumped-coding channel. There is first-order phase transition for low noise (left inset) and a second-order phase transition for high noise (right inset). **C**. Independent-coding channel with vanishing input noise. No phase transition is visible since the "bifurcation" happens in the limit of infinite output noise. **D**. The lumped-coding channel with vanishing input noise exhibits second-order phase transitions. From [1].

4.4.2 Critical exponents of the continuous phase transitions

Remark: The methods, results, figure, table, and text of this section (in particular, Eqs. 4.10-4.13, Tab. 4.1, and Fig. 4.10) are taken from a modified version of [1] which is currently under review. Content-wise, the methods, results, figure, table, and text of this section are identical to this modified version of [1]. However, here I have slightly modified some text for stylistic reasons and increased clarity.

Continuous phase transitions from different physical systems often behave very similarly around critical points (e.g. the Ising model of magnets at critical temperature or the liquid-gas transition at the critical point [89]). This phenomenon is known as *universality* and the universality class to which a system belongs can be characterized by critical exponents [89]. For example, the critical exponent β describes how the order parameter behaves for small temperature changes close to (but below) the critical temperature. In my system with mutual information and noise, β describes the behavior of threshold differences for noise values slightly smaller than critical noise values σ_c and R_c, i.e.

$$\Delta\theta \propto \left|\frac{\sigma - \sigma_c}{\sigma_c}\right|^{\beta_\sigma} \quad \text{for } \sigma < \sigma_c \tag{4.10}$$

and

$$\Delta\theta \propto \left(\frac{R - R_c}{R_c}\right)^{\beta_R} \quad \text{for } R > R_c \ , \tag{4.11}$$

respectively (Fig. 4.10A,B). I obtained critical exponents for both noise sources through fitting a monomial to the positive part of the threshold differences depending on one noise value while treating the other noise value as a parameter that I varied across 20 different values to get statistical robustness (Table 4.1). Similarly, I fitted the critical exponents for the eigenvalues which approach zero at critical noise values. Since the eigenvalues have finite values on both sides around the critical noise values, I separately fitted critical exponents for each side; e.g. for the output noise R, I fitted a monomial-function to both

$$|\lambda| \propto \left|\frac{\sigma - \sigma_c}{\sigma_c}\right|^{\phi_{R,l}} \quad \text{for } R < R_c \tag{4.12}$$

and

$$|\lambda| \propto \left(\frac{R - R_c}{R_c}\right)^{\phi_{R,r}} \quad \text{for } R > R_c \ , \tag{4.13}$$

again with 20 different values of input noise to get statistical robustness (Fig. 4.10C,D; Table 4.1). I found all the critical exponents for the eigenvalues to be approximately 1, while the thresholds differences as order parameters have a critical exponent of $\beta \approx 0.5$, the value predicted by the mean-field theory for all continuous phase transitions [118]. My results extend previous theoretical work which considered a population of two neurons with only input noise and already reported a critical exponent close to 0.5 [29]. For the critical exponent of the eigenvalues, ϕ, I have not found a resemblance in a statistical physics model.

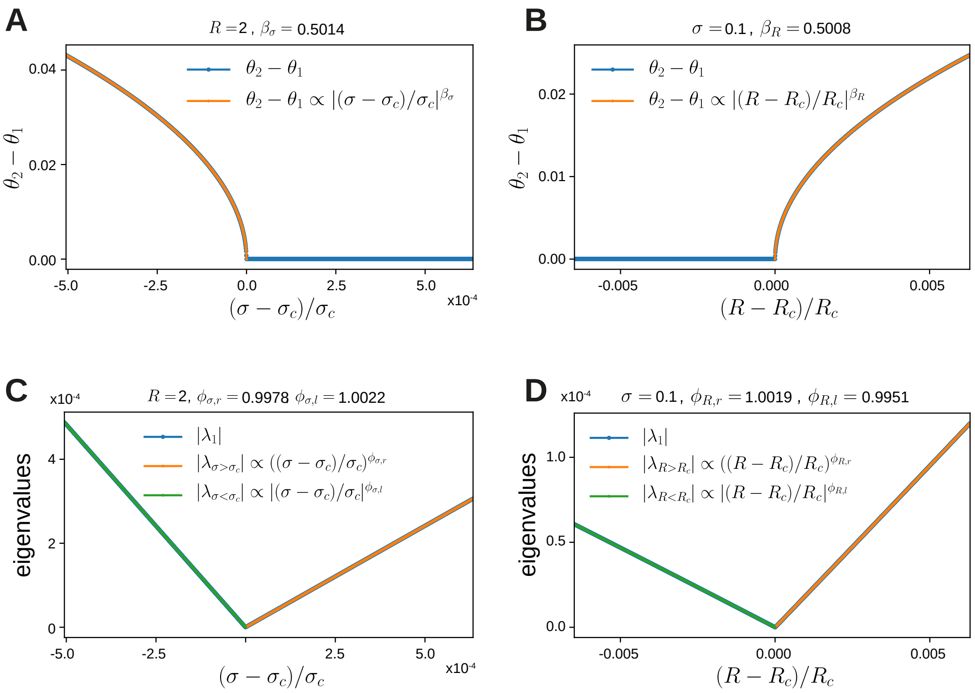


Figure 4.10: Obtaining critical exponents from fitting order parameter and eigenvalues in proximity of critical noise values of continuous phase transitions. **A**. Obtaining the critical exponent β_σ by fitting a monomial function to the threshold differences for input noise values slightly smaller than the critical input noise value σ_c. **B**. As in A, β_R is obtained by fitting output noise values slightly smaller than the critical output noise R_c. **C, D** Similarly one obtains the critical exponents of the eigenvalues of the Hessian of the information landscape, ϕ_l and ϕ_r, by fitting the eigenvalues for both slightly smaller (ϕ_l) and slightly larger (ϕ_r) noise values than the critical noise value. This figure is from an article that is based on [1] and that is currently under review.

4.4.3 The moment-generating function of the independent-coding channel

Remark: The methods, figure, and text of this section (in particular, Eqs. 4.14-4.17 and Fig. 4.11) are taken from a modified version of [1] which is currently under review. Content-wise, they are identical to this modified version of [1]. However, here I have slightly modified some text for stylistic reasons and increased clarity.

In statistical physics, the moment-generating function is directly related to the

Table 4.1: Critical exponents, their respective definitions and the values obtained from fitting. This table is from an article that is based on [1] and that is currently under review.

Crit. exp.	Definition	Fitted value (mean ± SEM)
β_σ	$\Delta\theta \propto \|(\sigma - \sigma_c)/\sigma_c\|^{\beta_\sigma}, \sigma < \sigma_c$	0.5027 ± 0.0018
β_R	$\Delta\theta \propto \|(R - R_c)/R_c\|^{\beta_R}, R > R_c$	0.5018 ± 0.0023
$\phi_{\sigma,l}$	$\|\lambda\| \propto \|(\sigma - \sigma_c)/\sigma_c\|^{\phi_{\sigma,r}}, \sigma < \sigma_c$	1.0034 ± 0.0019
$\phi_{\sigma,r}$	$\|\lambda\| \propto ((\sigma - \sigma_c)/\sigma_c)^{\phi_{\sigma,r}}, \sigma > \sigma_c$	0.9977 ± 0.0005
$\phi_{R,l}$	$\|\lambda\| \propto \|(R - R_c)/R_c\|^{\phi_{R,l}}, R < R_c$	0.9967 ± 0.0015
$\phi_{R,r}$	$\|\lambda\| \propto ((R - R_c)/R_c)^{\phi_{R,r}}, R > R_c$	1.0023 ± 0.0025

free energy of a system and plays a central role in studying its critical behavior and phase transitions [119] (see also App. A.1). The moment-generating function of an N-dimensional random variable $\vec{X}$ is:

$$M_{\vec{X}}(\vec{t}) = \langle e^{\vec{t}\cdot\vec{X}} \rangle, \qquad \vec{t} \in \mathbb{R}^N \ . \tag{4.14}$$

In my case, for the independent-coding channel and $N = 2$, the moment-generating function regarding the output variable is:

$$\begin{aligned} M_{\vec{k}}(\vec{t}) &= \langle e^{\vec{t}\cdot\vec{k}} \rangle & (4.15)\\ &= \sum_{k_1}\sum_{k_2} \vec{k}\, P(k_1, k_2) e^{t_1 k_1 + t_2 k_2} & (4.16) \end{aligned}$$

with

$$P(k_1, k_2) = \int_s \prod_{i=1}^{N} \sum_{\nu_i} P(k_i|\nu_i) P(\nu_i|s) P(s) ds, \tag{4.17}$$

where $P(k_i|\nu_i)$ follows the Poisson distribution and $P(\nu_i|s)$ is given by the threshold θ_i and the noise model (Sec. 3.1, Eqs. 3.3, 3.12). In statistical physics, phase transitions are characterized by a non-analytic behavior of the moment-generating function [120]. Since the moment-generating function is a sum of exponentials it should be non-analytic only for $N \to \infty$ [121]. In my work, I characterized phase transitions by non-analytic behavior of the maximized mutual information. That way, I found phase transitions for N finite and as small as two. Interestingly, the moment-generating function in my case is a smooth function of the thresholds and also – when the thresholds are not optimized but fixed – of both noises (Fig. 4.11A-C). However, in my case, the moment-generating function *does* become a non-analytic function of the noises when the *optimized* thresholds are used (Fig. 4.11D,E).

4.5 Summary and discussion

In this chapter, I quantified the information loss in the case of suboptimal thresholds. Using (local) curvature of the information landscape (Sec. 4.1) and different

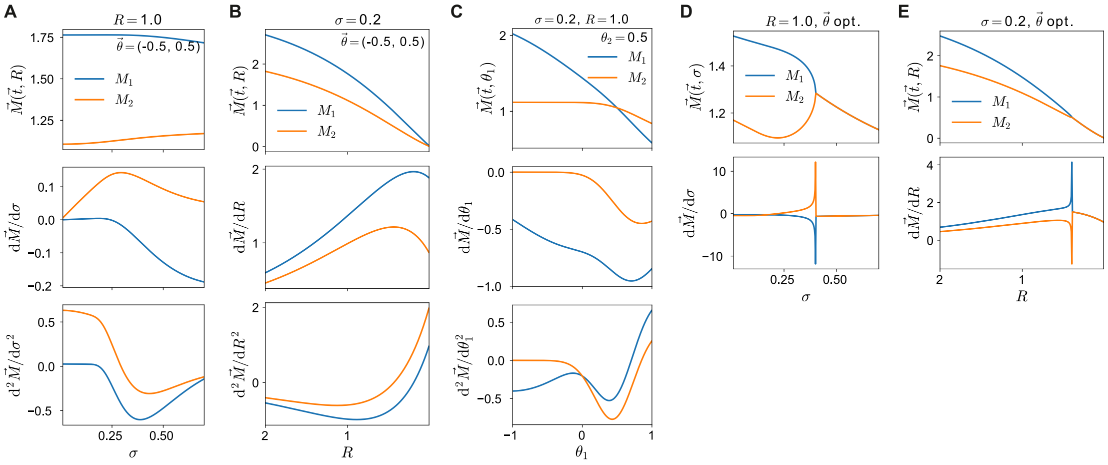


Figure 4.11: The moment generating function of the spike output vector $\vec{k}$ is smooth for fixed thresholds but not for optimized thresholds of both noise sources. **A**. The two components ($N = 2$) of the moment-generating function $\vec{M}(\vec{t})$ for $\vec{t} = (1, 1)$ depending on input noise σ. Output noise value and threshold vector are fixed to $R = 1$ and $\vec{\theta} = (-0.5, 0.5)$, respectively. The first two derivatives show no discontinuities. **B**. As A but depending on R with $\sigma = 0.2$. **C.** As A,B but depending on first threshold vector θ_1. **D, E.** As A,B but with optimized threshold vector for each noise value. The components of the moment-generating function show a bifurcation and the first derivatives show discontinuities. This figure is from an article that is based on [1] and that is currently under review.

ways of randomly sampling thresholds (Sec. 4.2) I found that, in general, the information loss relative to maximal information seems to decrease for larger populations. However, due to the difficulty of comparing distances across different dimensions, this quantification turned out to be not trivial. Moreover, different approaches lead to different quantitative results for different questions one wants to answer. During the process of quantifying the curvature of the information landscape I discovered that the landscape takes a particular shape at continuous threshold bifurcations: among at least one direction in the threshold space, the curvature of the landscape becomes zero meaning that perturbing thresholds in these directions leads to small information losses (Sec. 4.3). Finally, I was able to show that the threshold bifurcations respond to phase transitions: at critical noise levels, the system undergoes a first-order phase transition when thresholds bifurcate discontinuously and a second-order phase transition when thresholds bifurcate continuously (Sec. 4.4). To make a comprehensive comparison with physics' systems, I looked into critical exponents and the moment-generating function of the second-order phase transition.

4.5.1 Information loss due to suboptimal thresholds depends on measure

Systematically quantifying relative information loss due to suboptimal thresholds is not trivial since optimal thresholds can be perturbed in an N-dimensional space and quantification across dimensions is difficult. I tried two different approaches. First, through systematically quantifying particular directions in the threshold space along which I can perturb optimal thresholds, and second, randomly sampling thresholds. It turned out that the first approach is only feasible very locally around optimal thresholds (Fig. 4.2). There, it seems that the relative information loss is smaller for larger population size. However, the measure of perturbing thresholds by a fixed distance has a different impact depending on the dimension of the space: in high dimensional space it means that even though the length of the perturbation is the same, the perturbation in any single dimension is smaller. The second approach, randomly sampling thresholds, has the same caveat: the higher the dimension, the longer is the threshold vector on average when each component is sampled independently. Even more impactful, due to the law of large numbers, with increasing population size N, there is an increasingly strong shift away from very small or very large perturbations of the whole threshold vector towards mean perturbation values (Fig. 4.3).

I described three principle possibilities of randomly sampling threshold vectors. Two of them do not take into account that the average length of a sampled vector increases with its dimension N and thus cause small decreases of relative information loss with N (about 20% less loss for $N = 4$ compared to $N = 1$, Secs. 4.2.1 and 4.2.2). The third possibility makes sure that the average length of the sampled vector is constant across N. As a consequence each single threshold is only slightly perturbed for large N, thus causing a large decrease of information loss with N (about 80% less loss for $N = 4$ compared to $N = 1$, Sec. 4.2.1). It is not possible to give a clear answer on which measure is the right "one" and the measure to use depends on the concrete question one wants to answer. Concrete questions are, for example, how strong the evolutionary pressure is to optimize randomly distributed thresholds towards the optimum in general, or how deleterious mutations of already optimized thresholds are, or if the answer to the two previous questions depends on population size. Nevertheless, it is true to say that the importance of thresholds being precisely at the optimum decreases with N, although to different extents depending on the measure.

The conclusion is, that when maximizing information in future studies, instead of just finding optimal solutions, one also has to check how much worse suboptimal solutions actually are. For that, one also has to think about which measure to use when quantifying information loss. In particular, it is very hard to compare systems with different dimensions, i.e. different number of thresholds, with each other.

4.5.2 Threshold bifurcations resemble phase transitions

The phase transitions at critical noise levels in my work resemble phase transition from physics: maximizing the mutual information corresponds to minimizing free energy, noise corresponds to temperature, and the threshold differences correspond to order parameters. As in physical systems, the orders of my phase transitions are consistently linked to the continuity of the threshold differences: a continuous (discontinuous) order parameter corresponds to a second (first) order phase transition. At finite noise levels, the lumped-coding channel undergoes discontinuous threshold bifurcations which correspond to first-order phase transitions. In contrast, for the independent-coding channel, the threshold differences change continuously and the phase transitions are of second-order. My results suggest that input and output noise influence the mutual information in a very similar way that temperature or pressure affect free energy in physical systems [89]: both noise sources act as external parameters with respect to which the phase transition occurs (Figs. 4.7 and 4.9).

Critical exponents

For the case of continuous order parameters, i.e. optimal threshold differences of the independent-coding channel, I found critical exponents of the order parameter to be 0.5 – irrespective of the noise source (Fig. 4.10 and Tab. 4.1). This value corresponds to the mean-field theory of continuous phase transitions [118], which underscores the similarity of my phase transitions to those of physical systems. Furthermore, I determined the critical exponents for eigenvalues of the Hesse matrix of the information landscape and found them to be 1. Again the source of the noise has no impact which underscores the previously made statement, that additive input noise and Poisson output noise have a similar influence.

Discontinuity of the moment-generating function

In statistical physics, phase transitions occur where the moment-generating function of the state distribution shows non-analytic behavior, which only happens in the thermodynamic limit of having $N \to \infty$ many particles [90]. With my system of a neural population optimized for information encoding, the moment-generating function of the output distribution shows non-analytic behavior for $N = 2$ neurons (Fig. 4.11). However, this occurs only when using the optimized thresholds (which change with varying noise) but not when using fixed (even for varying noise) thresholds. The cause for this difference remains unclear, since I could only carry out numerical calculations. Thus, in contrast to models from statistical physics, I did not have access to analytic expressions which could help me understand why the non-analytic behavior occurs.

I provide an extensive comparison between bifurcations of the optimal thresholds and phase transitions studied in physics in the following chapter.

5 Discussion and outlook

Remark: Table 5.1 and some of the text in this chapter are part of an article entitled *Efficient population coding depends on stimulus convergence and source of noise* which has been written together with Shuai Shao and Julijana Gjorgjieva. The article has been uploaded to the preprint server *bioRxiv* [1] and a modified version of it is currently under review for publication in the journal *PLOS Computational Biology.* All content from that article which is part of this chapter was my contribution to the article unless specifically mentioned otherwise. The current rules of self-citation require me to strictly separate old from new content, but contrary to previous chapters, I have new and old material closely intermixed in this chapter. To not impair the typeface too much, I denote new content which has not previously been published by putting a dagger symbol, †, at its beginning and a double dagger symbol, ‡, at its end (e.g. †This sentence is new.‡).

†In this thesis, I have applied the efficient coding framework to a neural population with binary nonlinearities encoding a one-dimensional stimulus corrupted by two noise sources (additive input noise and output noise due to a stochastic spike generation process). I optimized the respective thresholds of the nonlinearities such that the mutual information between stimulus and population output is maximized. This way, I was able to characterize how the optimal threshold diversity depends on the two noise strengths. I did this for two different channel types, namely encoding the stimulus with the *lumped* output of the population or with its *independent* output. In summary, I have found that

- the lumped-coding channel encodes less information for all finite noise levels, meaning that lumping of the output is an additional form of noise (Fig. 3.2, Sec. 3.3).
- the optimal thresholds of the neurons' nonlinearities are all distinct at low noise and all equal at high noise and the transitions happen through subsequent bifurcations as a function of each noise (Fig. 3.4, Sec. 3.4).
- these bifurcations are continuous for the independent-coding channel and discontinuous for the lumped-coding channel (Fig. 3.4, Sec. 3.4).
- the two noise sources influence information encoding and optimal thresholds in a very similar way in that sense that each noise predicts fully distinct thresholds at lower noise levels and equivalent thresholds at higher noise levels,

with subsequent threshold bifurcations in between. Furthermore, the nature of the bifurcations does not depend on which noise source is varied but on the channel type (Fig. 3.4, Sec. 3.4).

- surprisingly, for some parameters, the optimal number of distinct thresholds behaves non-monotonously with noise, i.e. the number of optimal thresholds first decreases, then increases, and then decreases again with increasing noise (Fig. 3.12, Sec. 3.5).

- these results are robust with respect to the specific nature of the noise sources or the stimulus distribution (Secs. 3.6 and 3.7).

- the information landscape takes the form of a ridge at continuous threshold bifurcations, meaning that the information landscape is flat in certain directions and there is vanishing information loss when choosing suboptimal threshold combinations which lie on that ridge (Fig. 4.4, Sec. 4.3).

- the relative information loss achieved when the optimal thresholds are perturbed is smaller for larger neural populations (Fig. 4.3, Sec. 4.2).

- threshold bifurcations resemble phase transitions from statistical physics. As in physics, discontinuous threshold bifurcations resemble phase transitions of first-order, while continuous threshold bifurcations correspond to phase transitions of second-order (Fig. 4.7, Sec. 4.4). In general, the number of phase transitions increases with the population size.

In the following sections, I will discuss some of the results (Sec. 5.1), the assumptions of my model (Sec. 5.2), compare my model to other models from the literature (Tab. 5.1), and discuss the limitations (Sec. 5.3) and implications of my model (Sec. 5.4).[‡]

5.1 Discussion of the results

5.1.1 Lumping channels: trade-off between information loss and energy efficiency

Lumping the output of parallel channels into one effective channel causes a loss of information since lumping acts like a form of noise (Sec. 3.3). From an evolutionary perspective this appears to be counterproductive. So why would a biological system lump information transmission channels? A biological upside of combining information from multiple streams into one effective channel is the reduction of neurons needed for information transmission, thus potentially saving space and energy. For example, the optic nerve has a strong incentive to reduce its total diameter since it crosses through the retina and thus causes a blind spot. On the other hand, for a given constraint on space and energy, it is favorable to have many thin, low-rate axons over fewer thick, high-rate axons [122,123], thus arguing against convergence.

However, at least for the retina, an intermediate degree of convergence is probably the optimal solution. One would expect that this degree of convergence depends on the location at the retina. At the center of the retina, there is small convergence from photoreceptors to retinal ganglion cells compared to the periphery [124]. This implies that a higher visual acuity is achieved by increasing information transmission at the cost of energy and space. In contrast, there does not seem to be any convergence in the early auditory pathway: At the first stage of the neural signaling process, one inner hair cell diverges to 10 to 35 auditory nerve fibers [55]. This lack of convergence might be due to the fact that, contrary to the retina, there is no pressure of having a thin ganglion. A recent theoretical study suggests that convergence can compensate for the information loss due to a nonlinear tuning curve with a small number of output states [125].

I only treated the extreme cases of full convergence (where all neurons are lumped into a single channel) and no convergence (no lumping). In principle, different combinations of partial convergence, e.g. lumping three outputs into two channels, are also possible. Partial lumping is a common strategy in sensory systems with different levels of convergence [113]. Furthermore, I assumed no weighting of inputs during the lumping process. This is an oversimplification since in biology spikes from different presynaptic neurons could have a different impact on the membrane potential of the postsynaptic neuron depending on the synaptic connection strengths. These individual weights can also be optimized [86].

5.1.2 Optimal number of distinct thresholds as a function of noise

The number of distinct optimal thresholds decreases with increasing noise of either kind at critical noise levels by successive bifurcations of the optimal thresholds (Sec. 3.4). I mapped these characteristic bifurcations of the optimal thresholds at critical noise levels to phase transitions of different orders with order parameters being the threshold differences. At finite noise levels, the lumped-coding channel undergoes discontinuous threshold bifurcations which correspond to first-order phase transitions with respect to noise where the threshold differences are the order parameters. In contrast, for the independent-coding channel, the threshold differences change continuously and the phase transitions are of second-order.

Interestingly, for a range of noise parameters, I found non-monotonic changes in the number of distinct optimal thresholds with noise levels (Sec. 3.5). A similar non-monotonicity has also been reported under maximization of the Fisher information for neurons encoding sound direction [84]. †The reasons for this behavior remains elusive. Numerical imprecisions, however, can be excluded (Sec. 3.5.2). The non-monotonic threshold diversity seems to be related to the much more common non-monotonic threshold differences [62,85].‡ A related phenomenon in physics is that of *retrograde phenomena* [126]. For example, in a mixture of liquids, a phase transition from liquid to gas, followed by another transition from gas to liquid, and then liquid to gas again can be observed while increasing temperature [126].

5.1.3 Information loss at non-optimal thresholds

An important but often neglected question for optimal coding theories is how much worse suboptimal solutions are in comparison to optimal ones in terms of information transmission. In the independent-coding channel, near critical noise levels, the information landscape becomes flat in the directions of principal curvature (Sec. 4.3). This suggests that multiple threshold combinations yield nearly identical information, a property of the neuronal population that is closely related to the concept of “stiff vs sloppy modeling”, whereby a system’s output is insensitive to changes in “sloppy” directions of the parameter space, but very sensitive to changes in “stiff” directions [127–130]. Hence, even populations that utilize suboptimal thresholds often achieve information very close to the maximal, and it is unclear whether such small information differences could be measured experimentally. This also raises the question of whether a few percent more information about a stimulus realized by optimal codes could be sufficiently beneficial for the performance of a sensory system to become a driving force during evolution. †Therefore, I looked at the information loss when sampling thresholds randomly compared to the maximum information when using optimal thresholds and found that relative information loss is smaller the larger the neural population. However, the exact result depends on the question and the measure (Sec. 4.2). For example, the decrease of information loss with population size is smaller when each threshold component is independently perturbed compared to when the magnitude of the perturbation of the whole threshold vector is fixed across population sizes. The former case would resemble a mutation of a gene that influences the threshold of a single neuron, while the latter would resemble a mutation of a gene that influences the thresholds of all neurons.‡

It has been shown that mutations that have very small effects on evolutionary fitness are fixated in a population with a probability almost irrespective of the mutation being advantageous or deleterious [131, 132]. †This analysis demonstrates that despite the prominence and success of efficient coding frameworks that optimize information about a stimulus utilizing population codes with multiple neurons, one should be cautious when interpreting the optimal solutions. It turns out that even population codes that utilize suboptimal thresholds can achieve information very close to the optimally possible, putting into question whether differences in the biological implementations of such codes could even be detected.‡

On the other hand, in certain sensory systems like the retina, entire populations of retinal ganglion cells perform multiple functions [133–135] or fulfill different computations under different light conditions [136]. For such systems, there must be a fundamental trade-off in performance, since such a system cannot be optimal for all functions simultaneously [61, 137]. The sloppiness of nearly-equivalent optimal thresholds that I observed near critical noise levels should resolve when taking into account that neurons have multiple constraints and often perform more than just one function. †Incorporating several constraints into a model, however, requires additional assumptions about weighting constraints as they often have opposite effects: a metabolic constraint, for example, favors many thick axons with low firing rates, while a space constraint favors few thin axons with high firing rates [2].‡

5.1.4 Analogies and differences to phase transitions in statistical physics

†There are several direct analogies between phase transitions in this work and in statistical physics, the most prominent being: maximizing the mutual information corresponds to minimizing free energy, noise corresponds to temperature, the threshold differences correspond to order parameters, and a continuous behavior of the threshold differences corresponds to a second-order phase transition. However, there are also some noteworthy differences, for example, the critical exponent of the eigenvalue of the Hessian has not been reported in statistical physics, and my system has several order parameters. Since in statistical physics mostly second-order phase transitions are of interest [90], I extensively discuss the analogies and differences between statistical physics and my system using the independent-coding channel.‡

5.1.4.1 Symmetry break during phase transition

In statistical physics models, the transition of an order parameter from zero to non-zero values is accompanied by a symmetry break of the system. Examples are the symmetry break introduced by magnetization of a ferromagnetic material below the critical temperature,[1] by the non-miscibility of liquids below the critical temperature, and by the occurrence of hydrogen-bounds through the transition from vapor to liquid water [89]. Similarly, there is a symmetry break in my system as optimal thresholds become unequal at critical noise levels and thus the statistical equivalence of neurons breaks.

5.1.4.2 Critical exponents

I found critical exponents of the order parameter β to be 0.5, with respect to both input and output noise (Tab. 4.1, Sec. 4.4.2). This value corresponds to the mean-field theory of continuous phase transitions [118], which underscores the similarity of my phase transitions to those of physical systems. Since mean-field theory ignores statistical fluctuations, the measured exponents of physical systems are in most cases different from the ones predicted by theory and are referred to as

[1] Above the critical temperature there is no magnetization and the material is completely symmetric in all directions.

"anomalous" exponents [89]. In my model, the mutual information already takes into account statistical fluctuations, and appears to be an analytic function of the thresholds. Therefore, I do not expect an analogous mechanism that would lead to anomalous scaling exponents. Another critical exponent in my continuous phase transitions is the exponent of the smaller eigenvalue of the Hessian of the information landscape. I found this critical exponent to be 1. As before, the source of the noise has no impact on the critical exponent, which again highlights that additive input noise and Poisson output noise have a similar influence.

†In physical systems, it is usually computationally very expensive to calculate the Hessian of the free energy due to its high-dimensional landscape and thus can only be done by using symmetry arguments or other approximations to reduce the dimension of the landscape [138, 139]. It has been noted before in perturbation analysis that if one eigenvalue of the Hessian[2] goes to zero, the perturbations in the respective directions diverge which resembles a second-order phase transition. In any case, I have not found any evidence in the existing work that relate critical exponents to the eigenvalues of the Hessian.‡

5.1.4.3 Order parameters: optimized values vs. statistical quantities

In standard phase transitions, the order parameters are statistical quantities since they are the moment of a function, e.g. magnetization in the Ising model is the mean over spin directions [90]. My order parameters are not statistical variables but are obtained by optimizing mutual information. It is possible that they are related to statistical moments of some function of neural activity or to a function of statistical moments of neural activity, however, I have not found such a relationship.

5.1.4.4 More than one order parameter

In contrast to most physical systems, my system has more than one order parameter, specifically the number of subsequent threshold differences, i.e. the number of neurons minus one. My scenario with three neurons shows similarities with a system with three mixed liquids where the miscibility depends on the liquids' relative concentration differences [141]. As the temperature varies, the system undergoes two phase transitions at which the miscibility changes: from having one phase in which all three liquids are miscible, to two phases where in one phase two liquids are miscible but which is separated from a second phase containing the third liquid, to three phases where none of the liquids are miscible with each other. This corresponds to my system where the number of distinct thresholds varies with noise: from all three thresholds being distinct, to two being distinct, to all three being equal.

[2] To be precise: the Hessian of the Legendre transform of the logarithm of the moment-generating function [140].

5.1.4.5 Lumped-coding channel: second-order phase-transition in the limit of one noise approaching zero

†With input and output noise, the lumped-coding channel in general shows discontinuous bifurcations, i.e. first-order phase transitions. For one of these noise sources becoming smaller and approaching zero, however, the discontinuity becomes smaller and approaches a continuous bifurcation, i.e. second-order phase transition. This is comparable to the water-vapor phase transition with respect to temperature and pressure [89]: in general, the transition from water to vapor is a first-order phase transition with respect to both temperature and pressure. However, there exist a critical point of temperature and pressure where this phase transition becomes second-order.[3] As with the lumped channel and one noise approaching zero, the first-order phase transition from water to vapor continuously approaches a second-order phase transition when temperature and pressure approach the critical point.‡

5.2 Assumptions in my model and comparison to other theoretical frameworks

5.2.1 Assumptions about the stimulus

I considered the encoding of a static stimulus, even though natural stimuli have correlations in space and time. Previous studies have exploited their correlation structure to explain various aspects of sensory coding, for example, the size and shape of receptive fields of RGCs [38, 40, 41, 78, 81, 83]. Since correlations in the stimulus are thought to reduce effective noise values [78], by considering stimuli independent in time, I likely underestimated effective noise levels.

Moreover, my coding framework assumed a one-dimensional stimulus; thus, it is appropriate for explaining the number of the population's distinct thresholds which encode a *single* stimulus feature – this could be the contrast at a single spatial position on the retina (as found to be coded by two different types of retinal ganglion cells that encode the same linearly filtered stimulus [29]), or sound intensity at a single frequency (as found to be coded by ANFs, where many ANFs get input from the same inner hair cell [56,143]). †Throughout this study I investigated the encoding of a one-dimensional stimulus drawn from a symmetric distribution; however, natural stimulus distributions are skewed and thus asymmetric [41]. Since the relationship between stimulus distribution and input noise distribution is what matters in this respect, using a skewed stimulus distribution could potentially qualitatively effect the pattern of optimal thresholds.‡

[3] Namely at 647 K and 220 bar [142].

5.2.2 Assumptions about the noise

In my work, I studied how noise entering at different stages affects information encoding without going into detail about the origins of this noise. In the mammalian retina, multiple sources of noise can be identified in the retinal circuits, including from the photoreceptors [144–147] or at the bipolar cell output synapses [78, 148–150]. In the case that my model was applied to coding by RGCs at the same spatial locations and with the same visual feature, these sources would all count as input noise. Their relative contributions and the total magnitude could change with ambient light level, especially when one considers the signal-to-noise ratio [151].

The output noise originates after the thresholding nonlinearity. It can be linked to random fluctuations in the membrane potential (e.g. due to random openings and closings of ion channels [26]) or to stochastic vesicle release at synapses [152]. This noise is often taken to follow Poisson statistics where the variance in output scales with the output strength. [†]I have shown that the exact statistics of the stochastic spike generation process have no qualitative impact on the optimal encoding as long as the spontaneous rate is not affected. All these types of output noise which are based on constraining the firing rate are multiplicative output noises [76]. In principle, also additive output noise can be considered when the signal processing cascade after the spike generation process might introduce considerable quantities of noise [76]. Similarly, input noise in sensory systems can also be multiplicative, where the noise strength scales with the stimulus values [77].[‡]

5.2.3 Binary nonlinearities

I modeled each neuron in the population solely with a binary nonlinearity. This nonlinearity describes the tuning curve of the neuron as a function of a given stimulus feature. In general, a tuning curve with respect to a stimulus feature is measured by reverse correlating the stimulus variable with the output variable and fitting a linear-nonlinear model [14]. The linear part of the model denotes the stimulus feature to which the neuron responds and the nonlinear part represents the tuning curve. I did not incorporate the linear part in my model but rather assumed that the input to the nonlinearity is already linearly preprocessed because simultaneous optimization under different noise sources and stimulus convergence would be mathematically intractable. I chose binary nonlinearities as they are theoretically optimal under certain conditions of high (and biologically plausible) Poisson noise [76, 86, 106]. For example the steepness of the tuning curve of the H1 blowfly neuron increases with contrast, and for high contrast – which corresponds to low noise – the tuning curve is almost binary [44]. However, under conditions of non-negligible input noise, the optimal nonlinearity could be interpreted to acquire a finite slope thus making my analysis relevant also for continuous nonlinearities with a sigmoidal shape. This is consistent with neuronal recordings.

5.2.4 Maximum firing rate vs. mean firing rate constraint

†Maximizing mutual information needs some form of constraint since with arbitrary large firing rates channel capacity can always be reached with any noise strength. Throughout most of this work, I have quantified output noise by a variable that is the product of the maximum firing rate of the neurons and the coding window length.‡ Such a constraint is motivated by a biophysical limit of a neuron's firing rate and the biological reality of a short reaction time. Instead of constraining the maximum firing rate, one can also constrain the *mean* spike count [29, 62, 153], which would be interpreted as a metabolic constraint. †It can be implemented in different ways: a relatively simple way is to not optimize the value of one threshold but automatically assign it through the mean rate constrained [29] (Eq. 3.57, Sec. 3.6.3). A more involved way is that of optimizing all threshold values and an also optimizing an additional weight parameter that determines the optimal ratio of the firing rates among the neurons [62].‡ Maximum and mean rate constraints lead to qualitatively similar conclusions regarding the optimal number of thresholds (though the thresholds seem to be shifted to higher values), as shown in small populations of two neurons [29, 62].

5.2.5 Comparison to previous studies

Many previous studies make very similar assumptions but consider certain limiting scenarios, for instance considering only one noise source [29, 37, 85, 86], studying a population with only two neurons [29, 37, 76], introducing an additional source of additive output noise [76], or using different quantities to optimize [77, 84, 154]. Table 5.1 summarizes these studies with regards to the different optimization measures, constraints, information convergence strategies, sources of noise and neuronal population size. While my results are in agreement with these previous studies in the specific limiting conditions, I extended the optimal coding framework by mapping the full space of noise and stimulus convergence thus linking and extending previous findings.

5.3 Limitations of my framework

†In this section, I elaborate on the limits of my framework. This includes the question about how much computation is already performed in sensory organs in contrast to pure information transmission, and the fact that for most animals not all information is equally relevant from an evolutionary perspective.‡

5.3.1 Information transmission vs. computation

†In the retina, RGCs are the third layer of neurons after the photoreceptors and the bipolar cells, with amacrine cells in-between [49]. RGCs, in general, integrate inputs from many bipolar cells and amacrine cells in a nonlinear way, often performing very

Table 5.1: Comparison of different studies with regards to the different optimization measures, constraints, information convergence strategies, sources of noise and neuronal population size. MI stands for Mutual Information and MSE for Mean Square Error. Modified from [1].

Study	**Optimality measure**	**Constraint**	**Lumped or indep.**	**Input or output noise**	**# Neurons**
My work	MI	Max. rate	Both	Both	≤ 6
Brinkman et al. [76]	MI and MSE	Max. rate	Indep.	Both	2
Gjorgjieva et al., 2014 [62]	MI and MSE	Mean rate	Indep.	Both	2
Kastner et al. [29]	MI	Mean rate	Indep.	Input	2
Gjorgjieva et al., 2019 [37]	MI and MSE	Max. rate	Indep.	Output	any
Nikitin et al. [86]	MI	Max. rate	Lumped	Output	4
McDonnell et al. [85]	MI	Max. rate	Lumped	Input	≤ 15
Bethge et al. [59]	MI and MSE	Max. rate	Indep.	Output	4
Harper and McAlpine [84]	Fisher info	Bell-shaped tuning curves	Indep.	Output	200
†Wang et al., [153]	Fisher I., MI, L_p measure	"Meta-tuning curves"	Indep.	Output	any‡

complex computations, e.g. detecting direction-specific movements of objects [155], while for many RGCs it is still unknown which computations they perform [18]. Thus, it might be naive to assume that sensory neurons like RGCs solely transmit information about contrast levels of their receptive fields to the brain. Since my framework makes no explicit assumptions about the feature of the input stimulus in principle it could also be applied in the context of complex input features, e.g. object movements in specific directions. However, one assumption of my framework is that all preprocessing of the input is linear and this assumption is usually not justified for such complex stimulus features.‡

5.3.2 Not all information is equally important

†Contrary to a digital camera, the goal of sensory perception is not to faithfully represent the light intensity at every pixel of the retina but to encode as much information as possible about biologically *relevant* stimuli [155]. Relevant stimuli for

most species include information about danger, food, or communication with fellow animals. Most of these stimuli include some form of stimulus change, e.g. due to object movement. In principle, the importance of changing stimuli is incorporated in my framework, since it takes linearly filtered stimuli as input. A biphasic linear filter, for example, highlights changing stimuli while constant stimuli are repressed. Sensory organs in general receive feedback from the brain, which modulates stimulus processing. In the auditory system, there is extensive feedback from the cortex back to the basilar membrane of the cochlea where the transduction from sound signal to a neural signal happens [156]. In the visual system, the feedback to the retina seems to be much smaller and its details mostly remain elusive [157, 158]. With feedback from the brain to the sensory organs, however, higher-order mechanisms like attention can influence linear filters at the beginning or even before neural processing. That way, the information about biologically relevant stimuli might be very differently processed than non-relevant stimuli. However, their processing can still be modeled by an LNP model. Relevant and non-relevant stimuli might have a different (filtered) stimulus distribution which should be considered.‡

5.3.3 Optimal shape of nonlinearity

†Previous studies have optimized nonlinearities of neural information encoding under different assumptions and constraints [43,76,77,86,153]. In the limit of very low noise and no metabolic constraints, efficient coding predicts that the nonlinearity should be the cumulative distribution function of the input distribution [43, 77].[4] In my work, I assumed nonlinearities with a fixed shape – mathematically binary but effectively sigmoidal in the presence of input noise. Since I only optimized the thresholds of the nonlinearity, my tuning curves are in general suboptimal. For high output noise, however, it has been shown that a very steep or even binary nonlinearity (corresponding to small or zero input noise in my model) is optimal [76,86].‡

5.4 Implications of my model

My model consists of a population of neurons that codes for a one-dimensional stimulus. It is a general model that could apply to any sensory system, including the coding of sound intensity in auditory nerve fibers [56, 143], the coding of temperature in thermosensation by heat- and cold-activated ion channels [7, 159], the coding of vibration frequency by mechanosensory neurons [8, 160], and the coding of contrast by retinal ganglion cells coding for the same visual feature with different thresholds [29].

[4] The nonlinearity being the cumulative of the stimulus distribution maximizes output entropy since it “equalizes” the response distribution, i.e. all output values occur with equal probability. Intuitively speaking: the nonlinearity is most sensitive to stimulus values where it has its steepest part and this most sensitive part should be matched to the most probable stimuli.

†In work done by my colleague Shuai Shao, it was possible to show that the results of my framework are in agreement with data from ANFs [1]. However, the theory is also relevant for other sensory systems where each neuron of a population encodes the same stimulus feature and where the two noise sources can be separately measured. When verifying predictions from models using the efficient coding hypothesis, noise is usually inferred but not manipulated [1, 29, 72, 86]. However, since the level of the input noise is given as the ratio of the variances of the input noise distribution and the stimulus distribution [76], it could be possible to vary the noise level through varying the stimulus distribution. For example, reduced background illumination levels [44, 151] or reduced general sound intensities [161] should reduce the signal to noise ratio.‡

†The implications of the non-monotonous threshold behavior on biological systems should be insignificant since it should not make a difference for sensory systems if the optimal number of distinct thresholds only decreases with increasing noise or also increases for some parameters. Therefore, it is more a peculiarity that is not in accordance with previous results and the general assumption that optimal redundancy always increases with increasing noise.‡ A related phenomenon in physics is that of *retrograde phenomena* [126]. For example, in a mixture of liquids, a phase transition from liquid to gas, followed by another transition from gas to liquid, and then liquid to gas again can be observed while increasing temperature [126]. †Nevertheless, these non-monotonicities remain a peculiar and surprising phenomenon that deserves further theoretical investigation.‡

†In this thesis, I have investigated the coding of populations of sensory neurons assuming that sensory organs have optimized their function during evolution. I used computational modeling approaches to simply the problem and proposed a framework for how sensory populations support efficient information transmission to the brain. In contrast to previous work, I included different sources of sensory noise which corrupts information transmission. My results demonstrates that when there is significant amount of noise in the system, neurons in the sensory populations should diversify their functional properties in order to transmit optimal information about the sensory stimulus they encode. Therefore, my work makes predictions about coding strategies that can be applied to different sensory systems. Additionally, my work contributes to the understanding about the tradeoffs between information loss and signal compression due to signal convergence on information transmission.‡

Bibliography

[1] K. Röth, S. Shao, and J. Gjorgjieva, "Efficient population coding depends on stimulus convergence and source of noise," *bioRxiv*, 2020.

[2] P. Sterling and S. Laughlin, *Principles of neural design.* MIT Press, 2015.

[3] G. L. Fain, R. Hardie, and S. B. Laughlin, "Phototransduction and the evolution of photoreceptors," *Current Biology*, vol. 20, no. 3, pp. R114 – R124, 2010.

[4] C. Geisler, *From Sound to Synapse: Physiology of the Mammalian Ear.* Oxford University Press, 1998.

[5] G. Si, J. K. Kanwal, Y. Hu, C. J. Tabone, J. Baron, M. E. Berck *et al.*, "Structured odorant response patterns across a complete olfactory receptor neuron population," *Neuron*, vol. 101, no. 5, pp. 950–962.e7, 2019.

[6] A. Zimmerman, L. Bai, and D. D. Ginty, "The gentle touch receptors of mammalian skin," *Science*, vol. 346, no. 6212, pp. 950–954, 2014.

[7] A. P. Ajay Dhaka, Veena Viswanath, "Trp ion channels and temperature sensation," *Annual Review of Neuroscience*, vol. 29, pp. 135–161, 2006.

[8] V. Mountcastle, M. Steinmetz, and R. Romo, "Frequency discrimination in the sense of flutter: psychophysical measurements correlated with postcentral events in behaving monkeys," *Journal of Neuroscience*, vol. 10, no. 9, pp. 3032–3044, 1990.

[9] C. C. Bell, "Mormyromast electroreceptor organs and their afferent fibers in mormyrid fish. iii. physiological differences between two morphological types of fibers," *Journal of Neurophysiology*, vol. 63, no. 2, pp. 319–332, 1990.

[10] J. L. Kirschvink, M. M. Walker, and C. E. Diebel, "Magnetite-based magnetoreception," *Current Opinion in Neurobiology*, vol. 11, no. 4, pp. 462 – 467, 2001.

[11] G. E. Pugh, *The biological origin of human values.* Basic Books, New York, 1977.

[12] C. U. M. Smith, *Biology of sensory systems.* John Wiley & Sons, 2008.

[13] X. Pitkow and M. Meister, *The Cognitive Neurosciences V.* MIT Press, 2014, ch. Neural computation in sensory systems.

[14] P. Dayan and L. F. Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems.* MIT Press, 2001.

[15] B. Seybold, E. Phillips, C. Schreiner, and A. Hasenstaub, "Inhibitory actions unified by network integration," *Neuron*, vol. 87, no. 6, pp. 1181 – 1192, 2015.

[16] M. Yazdi and T. Bouwmans, "New trends on moving object detection in video images captured by a moving camera: A survey," *Computer Science Review*, vol. 28, pp. 157 – 177, 2018.

[17] T. Baden, T. Euler, and P. Berens, "Understanding the retinal basis of vision across species," *Nature Reviews Neuroscience*, vol. 21, no. 1, pp. 5–20, 2020.

[18] T. Baden, P. Berens, K. Franke, M. R. Rosón, M. Bethge, and T. Euler, "The functional diversity of retinal ganglion cells in the mouse," *Nature*, vol. 529, no. 7586, pp. 345–350, 2016.

[19] M. Vater and M. Kössl, "Comparative aspects of cochlear functional organization in mammals," *Hearing Research*, vol. 273, no. 1, pp. 89 – 99, 2011, comparative Studies of the Ear.

[20] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[21] D. Attwell and S. B. Laughlin, "An energy budget for signaling in the grey matter of the brain," *Journal of Cerebral Blood Flow & Metabolism*, vol. 21, no. 10, pp. 1133–1145, 2001.

[22] H. B. Barlow, *Possible principles underlying the transformations of sensory messages.* MIT press, 1961, ch. 13, p. 217–234.

[23] J. J. Atick, "Could information theory provide an ecological theory of sensory processing?" *Network: Computation in Neural Systems*, vol. 3, no. 2, pp. 213–251, 1992.

[24] R. C. S. Wong, S. L. Cloherty, M. R. Ibbotson, and B. J. O'Brien, "Intrinsic physiological properties of rat retinal ganglion cells with a comparative analysis," *Journal of Neurophysiology*, vol. 108, no. 7, pp. 2008–2023, 2012.

[25] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.

[26] A. A. Faisal, L. P. J. Selen, and D. M. Wolpert, "Noise in the nervous system," *Nature Reviews Neuroscience*, vol. 9, pp. 292–303, 2008.

[27] F. Naarendorp, T. M. Esdaille, S. M. Banden, J. Andrews-Labenski, O. P. Gross, and E. N. Pugh, "Dark light, rod saturation, and the absolute and incremental sensitivity of mouse cone vision," *Journal of Neuroscience*, vol. 30, no. 37, pp. 12 495–12 507, 2010.

[28] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, no. 9, pp. 1875–1902, Sep. 2005.

[29] D. B. Kastner, S. A. Baccus, and T. O. Sharpee, "Critical and maximally informative encoding between neural populations in the retina," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, pp. 2533–2538, 2015.

[30] G. J. Chader, J. Weiland, and M. S. Humayun, "Artificial vision: needs, functioning, and testing of a retinal electronic prosthesis," in *Neurotherapy: Progress in Restorative Neuroscience and Neurology*, ser. Progress in Brain Research. Elsevier, 2009, vol. 175, pp. 317–332.

[31] C. Lane, K. Zimmerman, S. Agrawal, and L. Parnes, "Cochlear implant failures and reimplantation: A 30-year analysis and literature review," *The Laryngoscope*, vol. 130, no. 3, pp. 782–789, 2020.

[32] H. P. N. Scholl, R. W. Strauss, M. S. Singh, D. Dalkara, B. Roska, S. Picaud, and J.-A. Sahel, "Emerging therapies for inherited retinal degeneration," *Science Translational Medicine*, vol. 8, no. 368, 2016.

[33] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of Physiology*, vol. 117, no. 4, pp. 500–544, 1952.

[34] B. Roska and M. Meister, "The retina dissects the visual scene into distinct features," in *The new visual neurosciences.* MIT Press, 2014, pp. 163–182.

[35] E. A. Hallem and J. R. Carlson, "Coding of odors by a receptor repertoire," *Cell*, vol. 125, no. 1, pp. 143–160, 2006.

[36] J. Tan, A. Savigner, M. Ma, and M. Luo, "Odor information processing by the olfactory bulb analyzed in gene-targeted mice," *Neuron*, vol. 65, no. 6, pp. 912–926, 2010.

[37] J. Gjorgjieva, M. Meister, and H. Sompolinsky, "Functional diversity among sensory neurons from efficient coding principles." *PLOS Computational Biology*, vol. 15, no. 11, pp. 1–38, 2019.

[38] J. J. Atick and A. N. Redlich, "Towards a theory of early visual processing," *Neural Computation*, vol. 2, no. 3, pp. 308–320, 1990.

[39] J. Atick and A. N. Redlich, "Convergent algorithm for sensory receptive field development," *Neural Computation*, vol. 5, no. 1, pp. 45–60, 1993.

[40] M. Haft and J. L. van Hemmen, "Theory and implementation of infomax filters for the retina," *Network: Computation in Neural Systems*, vol. 9, no. 1, pp. 39–71, 1998.

[41] C. P. Ratliff, B. G. Borghuis, Y.-H. Kao, P. Sterling, and V. Balasubramanian, "Retina is structured to process an excess of darkness in natural scenes," *Proceedings of the National Academy of Sciences*, vol. 107, no. 40, pp. 17 368–17 373, 2010.

[42] E. Doi, J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner *et al.*, "Efficient coding of spatial information in the primate retina," *Journal of Neuroscience*, vol. 32, no. 46, pp. 16 256–16 264, 2012.

[43] S. Laughlin, "A simple coding procedure enhances a neuron's information capacity," *Zeitschrift für Naturforschung c*, vol. 36, no. 9-10, pp. 910–912, 1981.

[44] N. Brenner, W. Bialek, and R. R. van Steveninck, "Adaptive rescaling maximizes information transmission," *Neuron*, vol. 26, no. 3, pp. 695–702, 2000.

[45] I. M. Winter, D. Robertson, and G. K. Yates, "Diversity of characteristic frequency rate-intensity functions in guinea pig auditory nerve fibres," *Hearing Research*, vol. 45, no. 3, pp. 191–202, 1990.

[46] G. S. Corrado, L. P. Sugrue, H. S. Seung, and W. T. Newsome, "Linear-nonlinear-poisson models of primate choice dynamics," *Journal of the Experimental Analysis of Behavior*, vol. 84, no. 3, pp. 581–617, 2005.

[47] F. Gabbiani and S. J. Cox, *Mathematics for neuroscientists.* Academic Press, 2017.

[48] R. H. Masland, "Neuronal diversity in the retina," *Current Opinion in Neurobiology*, vol. 11, no. 4, pp. 431–436, 2001.

[49] ——, "The neuronal organization of the retina," *Neuron*, vol. 76, no. 2, pp. 266–280, 2012.

[50] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, "Connectomic reconstruction of the inner plexiform layer in the mouse retina," *Nature*, vol. 500, no. 7461, pp. 168–174, 2013.

[51] J. R. Sanes and R. H. Masland, "The types of retinal ganglion cells: Current status and implications for neuronal classification," *Annual Review of Neuroscience*, vol. 38, no. 1, pp. 221–246, 2015.

[52] D. B. Kastner and S. A. Baccus, "Coordinated dynamic encoding in the retina using opposing forms of plasticity," *Nature Neuroscience*, vol. 14, no. 10, pp. 1317–1322, 2011.

[53] R. Segev, J. Puchalla, and M. J. Berry, "Functional organization of ganglion cells in the salamander retina," *Journal of Neurophysiology*, vol. 95, no. 4, pp. 2277–2292, 2006.

[54] Q. Cui, C. Ren, P. Sollars, G. Pickard, and K.-F. So, “The injury resistant ability of melanopsin-expressing intrinsically photosensitive retinal ganglion cells,” *Neuroscience*, vol. 284C, pp. 845–853, 11 2014.

[55] B. A. Nayagam, M. A. Muniak, and D. K. Ryugo, “The spiral ganglion: Connecting the peripheral and central auditory systems,” *Hearing Research*, vol. 278, no. 1, pp. 2–20, 2011.

[56] A. M. Taberner and M. C. Liberman, “Response properties of single auditory nerve fibers in the mouse,” *Journal of Neurophysiology*, vol. 93, no. 1, pp. 557–569, 2005.

[57] M. Tsunozaki and D. M. Bautista, “Mammalian somatosensory mechanotransduction,” *Current Opinion in Neurobiology*, vol. 19, no. 4, pp. 362–369, 2009.

[58] A. Borst and F. E. Theunissen, “Information theory and neural coding,” *Nature Neuroscience*, vol. 2, no. 11, pp. 947–957, 1999.

[59] M. Bethge, D. Rotermund, and K. Pawelzik, “Optimal neural rate coding leads to bimodal firing rate distributions,” *Network: Computation in Neural Systems*, vol. 14, no. 2, pp. 303–319, 2003.

[60] X.-X. Wei and A. A. Stocker, “Mutual information, fisher information, and efficient coding,” *Neural Computation*, vol. 28, no. 2, pp. 305–326, 2016.

[61] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo *et al.*, “Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space,” *Science*, 2012.

[62] J. Gjorgjieva, H. Sompolinsky, and M. Meister, “Benefits of pathway splitting in sensory coding,” *The Journal of Neuroscience*, vol. 34, no. 36, pp. 12 127–12 144, 2014.

[63] S. Ocko, J. Lindsey, S. Ganguli, and S. Deny, “The emergence of multiple retinal cell types through efficient coding of natural movies,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018, pp. 9389–9400.

[64] F. Soto, J.-C. Hsiang, R. Rajagopal, K. Piggott, G. J. Harocopos, S. M. Couch *et al.*, “Efficient coding by midget and parasol ganglion cells in the human retina,” *Neuron*, vol. 107, no. 4, pp. 656 – 666.e5, 2020.

[65] F. Rieke, D. A. Bodnar, and W. Bialek, “Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents,” *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 262, no. 1365, pp. 259–265, 1995.

[66] I. Dean, N. S. Harper, and D. McAlpine, “Neural population coding of sound level adapts to stimulus statistics,” *Nat Neurosci*, vol. 8, no. 12, pp. 1684–1689, Dec. 2005.

[67] J. A. White, J. T. Rubinstein, and A. R. Kay, “Channel noise in neurons,” *Trends in Neurosciences*, vol. 23, no. 3, pp. 131 – 137, 2000.

[68] P. Lillywhite and S. Laughlin, “Transducer noise in a photoreceptor,” *Nature*, vol. 277, no. 5697, pp. 569–572, 1979.

[69] W. Bialek, “Physical limits to sensation and perception,” *Annual review of biophysics and biophysical chemistry*, vol. 16, no. 1, pp. 455–478, 1987.

[70] P. Sterling, “How retinal circuits optimize the transfer of visual information,” in *The visual neurosciences*, J. S. Werner and L. M. Chalupa, Eds. MA: MIT Press, 2004, ch. 17, pp. 234–259.

[71] D. K. Warland, P. Reinagel, and M. Meister, “Decoding visual information from a population of retinal ganglion cells,” *Journal of Neurophysiology*, vol. 78, no. 5, pp. 2336–2350, 1997.

[72] E. J. Chichilnisky and F. Rieke, “Detection sensitivity and temporal resolution of visual signals near absolute threshold in the salamander retina,” *Journal of Neuroscience*, vol. 25, no. 2, pp. 318–330, 2005.

[73] G. Tkačik, C. G. Callan, and W. Bialek, “Information flow and optimization in transcriptional regulation,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 34, pp. 12 265–12 270, 2008.

[74] G. Tkačik, A. M. Walczak, and W. Bialek, “Optimizing information flow in small genetic networks,” *Physical Review E*, vol. 80, p. 031920, 2009.

[75] P. Berens, A. S. Ecker, S. Gerwinn, A. S. Tolias, M. Bethge, and W. S. Geisler, “Reassessing optimal neural population codes with neurometric functions,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 11, pp. 4423–4428, 2011.

[76] B. A. W. Brinkman, A. I. Weber, F. Rieke, and E. Shea-Brown, “How do efficient coding strategies depend on origins of noise in neural circuits?” *PLOS Computational Biology*, vol. 12, no. 10, pp. 1–34, 2016.

[77] Z. Wang, A. A. Stocker, and D. D. Lee, “Efficient neural codes that minimize Lp reconstruction error,” *Neural Computation*, vol. 28, no. 12, pp. 2656–2686, 2016.

[78] B. G. Borghuis, C. P. Ratliff, R. G. Smith, P. Sterling, and V. Balasubramanian, “Design of a neuronal array,” *Journal of Neuroscience*, vol. 28, no. 12, pp. 3178–3189, 2008.

[79] H. Barlow, “Redundancy reduction revisited,” *Network: Computation in Neural Systems*, vol. 12, no. 3, pp. 241–253, 2001.

[80] J. H. van Hateren, “A theory of maximizing sensory information,” *Biological Cybernetics*, vol. 68, no. 1, pp. 23–29, 1992.

[81] ——, “Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation,” *Journal of Comparative Physiology A*, vol. 171, no. 2, pp. 157–170, 1992.

[82] G. Tkačik, J. S. Prentice, V. Balasubramanian, and E. Schneidman, “Optimal population coding by noisy spiking neurons,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 32, pp. 14 419–14 424, 2010.

[83] Y. Karklin and E. P. Simoncelli, “Efficient coding of natural images with a population of noisy linear-nonlinear neurons,” in *Advances in neural information processing systems.* MIT Press, 2011, pp. 999–1007.

[84] N. S. Harper and D. McAlpine, “Optimal neural population coding of an auditory spatial cue,” *Nature*, vol. 430, pp. 682–686, 2004.

[85] M. D. McDonnell, N. G. Stocks, C. E. Pearce, and D. Abbott, “Optimal information transmission in nonlinear arrays through suprathreshold stochastic resonance,” *Physics Letters A*, vol. 352, no. 3, pp. 183–189, 2006.

[86] A. P. Nikitin, N. G. Stocks, R. P. Morse, and M. D. McDonnell, “Neural population coding is optimized by discrete tuning curves,” *Physical Review Letters*, vol. 103, p. 138101, 2009.

[87] X. Pitkow and M. Meister, “Decorrelation and efficient coding by retinal ganglion cells,” *Nature Neuroscience*, vol. 15, no. 4, pp. 628–635, 2012.

[88] C. Behrens, T. Schubert, S. Haverkamp, T. Euler, and P. Berens, “Connectivity map of bipolar cells and photoreceptors in the mouse retina,” *eLife*, vol. 5, p. e20041, 2016.

[89] H. E. Stanley, *Introduction to phase transitions and critical phenomena.* Oxford University Press, 1971.

[90] W. Greiner, L. Neise, and H. Stöcker, *Thermodynamics and statistical mechanics.* Springer Science & Business Media, 2012.

[91] E. Paul, “Phasenumwandlungen im ueblichen und erweiterten sinn, klassifiziert nach dem entsprechenden singularitaeten des thermodynamischen potentiales,” *Verhandlingen der Koninklijke Akademie van Wetenschappen (Amsterdam)*, vol. 36, pp. 153–157, 1933.

[92] M. B. Sachs, R. L. Winslow, and B. H. Sokolowski, "A computational model for rate-level functions from cat auditory-nerve fibers," *Hearing research*, vol. 41, no. 1, pp. 61–69, 1989.

[93] M. D. McDonnell, D. Abbott, and C. E. Pearce, "An analysis of noise enhanced information transmission in an array of comparators," *Microelectronics Journal*, vol. 33, no. 12, pp. 1079–1089, 2002.

[94] S. Linge and H. Petter Langtangen, *Programming for Computations-Python: A Gentle Introduction to Numerical Simulations with Python.* Springer Nature, 2016.

[95] E. Smith, *Python, the Fundamentals.* Cham: Springer International Publishing, 2020, pp. 19–50.

[96] M. J. Kochenderfer and T. A. Wheeler, *Algorithms for optimization.* Mit Press, 2019.

[97] F. Gao and L. Han, "Implementing the Nelder-Mead simplex algorithm with adaptive parameters," *Computational Optimization and Applications*, vol. 51, no. 1, pp. 259–277, 2012.

[98] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.

[99] V. J. Uzzell and E. J. Chichilnisky, "Precision of spike trains in primate retinal ganglion cells," *Journal of Neurophysiology*, vol. 92, no. 2, pp. 780–789, 2004.

[100] B. O'neill, *Elementary differential geometry.* Academic press, 2014.

[101] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[102] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau *et al.*, "Array programming with NumPy," *Nature*, vol. 585, p. 357–362, 2020.

[103] J. B. Ramsey, H. J. Newton, and J. L. Harvill, *The elements of statistics: With applications to economics and the social sciences.* Duxbury/Thomson Learning, 2002.

[104] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

[105] N. Balakrishnan, *Handbook of the logistic distribution.* CRC Press, 1991.

[106] S. Shamai, "Capacity of a pulse amplitude modulated direct detection photon channel," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 137, pp. 424–430(6), 1990.

[107] U. Fano, "Ionization Yield of Radiations. II. The Fluctuations of the Number of Ions," *Physical Review*, vol. 72, no. 1, pp. 26–29, Jul. 1947.

[108] N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate discrete distributions.* John Wiley & Sons, 2005, vol. 444.

[109] S. Nadarajah, "Useful moment and cdf formulations for the com–poisson distribution," *Statistical papers*, vol. 50, no. 3, pp. 617–622, 2009.

[110] M. Abramowitz, I. A. Stegun, and R. H. Romer, "Handbook of mathematical functions with formulas, graphs, and mathematical tables," 1988.

[111] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[112] D. J. Field, "What is the goal of sensory coding?" *Neural Computation*, vol. 6, no. 4, pp. 559–601, 1994.

[113] T. Euler, S. Haverkamp, T. Schubert, and T. Baden, "Retinal bipolar cells: elementary building blocks of vision," *Nature Reviews. Neuroscience*, vol. 15, no. 8, pp. 507–519, 2014.

[114] K. Crane, "A quick and dirty introduction to the curvature of surfaces," http://wordpress.discretization.de/geometryprocessingandapplicationsws19/a-quick-and-dirty-introduction-to-the-curvature-of-surfaces/, 2019.

[115] S. Dineen, *Multivariate calculus and geometry.* Springer, 2014.

[116] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2017.

[117] M. E. Muller, "A note on a method for generating points uniformly on N-dimensional spheres," *Commun. ACM*, vol. 2, no. 4, pp. 19–20, 1959.

[118] L. D. Landau, E. M. Lifšic, E. M. Lifshitz, and L. Pitaevskii, *Statistical physics: theory of the condensed state.* Butterworth-Heinemann, 1980, vol. 8.

[119] M. Bulmer, "Statistical inference," in *Principles of statistics.* Dover Publications, 1979, pp. 165–187.

[120] O. C. Martin, R. Monasson, and R. Zecchina, "Statistical mechanics methods and phase transitions in optimization problems," *Theoretical Computer Science*, vol. 265, no. 1, pp. 3–67, 2001.

[121] C.-N. Yang and T.-D. Lee, "Statistical theory of equations of state and phase transitions. i. theory of condensation," *Physical Review*, vol. 87, no. 3, p. 404, 1952.

[122] J. A. Perge, K. Koch, R. Miller, P. Sterling, and V. Balasubramanian, "How the optic nerve allocates space, energy capacity, and information," *Journal of Neuroscience*, vol. 29, no. 24, pp. 7917–7928, 2009.

[123] V. Balasubramanian and P. Sterling, "Receptive fields and functional architecture in the retina," *The Journal of Physiology*, vol. 587, no. 12, pp. 2753–2767, 2009.

[124] H. Kolb, "How the retina works: Much of the construction of an image takes place in the retina itself through the use of specialized neural circuits," *American Scientist*, vol. 91, no. 1, pp. 28–35, 2003.

[125] G. J. Gutierrez, F. Rieke, and E. Shea-Brown, "Nonlinear convergence preserves information," *bioRxiv*, 2019.

[126] A. Danesh, *PVT and phase behaviour of petroleum reservoir fluids.* Elsevier, 1998, vol. 47.

[127] K. S. Brown and J. P. Sethna, "Statistical mechanical approaches to models with many poorly known parameters," *Physical Review E*, vol. 68, p. 021904, 2003.

[128] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, "Universally sloppy parameter sensitivities in systems biology models," *PLOS Computational Biology*, vol. 3, no. 10, pp. 1–8, 2007.

[129] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, "Parameter space compression underlies emergent theories and predictive models," *Science*, vol. 342, no. 6158, pp. 604–607, 2013.

[130] D. Panas, H. Amin, A. Maccione, O. Muthmann, M. van Rossum, L. Berdondini, and M. H. Hennig, "Sloppiness in spontaneously active neuronal networks," *Journal of Neuroscience*, vol. 35, no. 22, pp. 8480–8492, 2015.

[131] M. Kimura, "Some problems of stochastic processes in genetics," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 882–901, 1957.

[132] ——, "Evolutionary rate at the molecular level," *Nature*, vol. 217, no. 5129, pp. 624–626, 1968.

[133] A. L. Fairhall, C. A. Burlingame, R. A. H. R. H. Narasimhan, J. L. Puchalla, and M. J. Berry, "Selectivity for multiple stimulus features in retinal ganglion cells," *Journal of Neurophysiology*, vol. 96, no. 5, pp. 2724–2738, 2006.

[134] T. A. Münch, R. A. da Silveira, S. Siegert, T. J. Viney, G. B. Awatramani, and B. Roska, "Approach sensitivity in the retina processed by a multifunctional neural circuit," *Nature Neuroscience*, vol. 12, 2009.

[135] S. Deny, U. Ferrari, E. Macé, P. Yger, R. Caplette, S. Picaud *et al.*, "Multiplexed computations in retinal ganglion cells of a single type," *Nature Communications*, vol. 8, no. 1, p. 1964, 2017.

[136] A. Tikidji-Hamburyan, K. Reinhard, H. Seitter, A. Hovhannisyan, C. A. Procyk, A. E. Allen *et al.*, "Retinal output changes qualitatively with every change in ambient illuminance," *Nature Neuroscience*, vol. 18, no. 1, pp. 66–74, 2015.

[137] W. F. Młynarski and A. M. Hermundstad, "Adaptive coding for dynamic sensory inference," *eLife*, vol. 7, p. e32055, 2018.

[138] R. Kikuchi, "Second hessian determinant as the criterion for order (first or second) of phase transition," *Physica A: Statistical Mechanics and its Applications*, vol. 142, no. 1, pp. 321 – 341, 1987.

[139] R. Bianco, I. Errea, L. Paulatto, M. Calandra, and F. Mauri, "Second-order structural phase transitions, free energy curvature, and temperature-dependent anharmonic phonons in the self-consistent harmonic approximation: Theory and stochastic implementation," *Phys. Rev. B*, vol. 96, p. 014111, Jul 2017.

[140] M. Helias and D. Dahmen, "Statistical field theory for neural networks," *arXiv preprint arXiv:1901.10416*, 2019.

[141] F. C. Campbell, *Phase diagrams: understanding the basics.* ASM International, 2012.

[142] M. D. Koretsky, *Engineering and chemical thermodynamics.* John Wiley & Sons, 2012.

[143] M. Liberman, "Single-neuron labeling in the cat auditory nerve," *Science*, vol. 216, no. 4551, pp. 1239–1241, 1982.

[144] D. M. Schneeweis and J. L. Schnapf, "The photovoltage of macaque cone photoreceptors: Adaptation, noise, and kinetics," *Journal of Neuroscience*, vol. 19, no. 4, pp. 1203–1216, 1999.

[145] P. Ala-Laurila, M. Greschner, E. J. Chichilnisky, and F. Rieke, "Cone photoreceptor contributions to noise and correlations in the retinal output," *Nature Neuroscience*, vol. 14, no. 10, pp. 1309–1316, 2011.

[146] J. M. Angueyra and F. Rieke, "Origin and effect of phototransduction noise in primate cone photoreceptors," *Nature Neuroscience*, vol. 16, no. 11, pp. 1692–1700, 2013.

[147] B. C. Hansen, D. J. Field, M. R. Greene, C. Olson, and V. Miskovic, “Towards a state-space geometry of neural responses to natural scenes: A steady-state approach,” *NeuroImage*, vol. 201, p. 116027, 2019.

[148] M. A. Freed, “Parallel cone bipolar pathways to a ganglion cell use different rates and amplitudes of quantal excitation,” *The Journal of Neuroscience*, vol. 20, no. 11, pp. 3956–3963, 2000.

[149] F. A. Dunn and F. Rieke, “The impact of photoreceptor noise on retinal gain controls,” *Current Opinion in Neurobiology*, vol. 16, no. 4, pp. 363–370, 2006.

[150] M. A. Freed and Z. Liang, “Synaptic noise is an information bottleneck in the inner retina during dynamic visual stimulation,” *The Journal of Physiology*, vol. 592, no. 4, pp. 635–651, 2014.

[151] K. Farrow, M. Teixeira, T. Szikra, T. J. Viney, K. Balint, K. Yonehara, and B. Roska, “Ambient illumination toggles a neuronal circuit switch in the retina and visual perception at cone threshold,” *Neuron*, vol. 78, no. 2, pp. 325–338, 2013.

[152] G. Murphy and F. Rieke, “Signals and noise in an inhibitory interneuron diverge to control activity in nearby retinal ganglion cells,” *Nature Neuroscience*, vol. 11, p. 318–326, 2008.

[153] Z. Wang, X.-X. Wei, A. A. Stocker, and D. D. Lee, “Efficient neural codes under metabolic constraints,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016, pp. 4619–4627.

[154] M. Bethge, D. Rotermund, and K. Pawelzik, “Optimal short-term population coding: When fisher information fails,” *Neural Computation*, vol. 14, no. 10, pp. 2317–2351, 2002.

[155] T. Gollisch and M. Meister, “Eye smarter than scientists believed: Neural computations in circuits of the retina,” *Neuron*, vol. 65, no. 2, pp. 150–164, 2010.

[156] P. H. Delano and A. B. Elgoyhen, “Editorial: Auditory efferent system: New insights from cortex to cochlea,” *Frontiers in Systems Neuroscience*, vol. 10, p. 50, 2016.

[157] J. Repérant, M. Médina, R. Ward, D. Miceli, N. Kenigfest, J. Rio, and N. Vesselkin, “The evolution of the centrifugal visual system of vertebrates. a cladistic analysis and new hypotheses,” *Brain Research Reviews*, vol. 53, no. 1, pp. 161 – 197, 2007.

[158] S. Schröder, N. A. Steinmetz, M. Krumin, M. Pachitariu, M. Rizzi, L. Lagnado *et al.*, “Arousal modulates retinal output,” *Neuron*, vol. 107, no. 3, pp. 487 – 495.e9, 2020.

[159] E. A. Lumpkin and M. J. Caterina, "Mechanisms of sensory transduction in the skin," *Nature*, vol. 445, no. 7130, pp. 858–865, 2007.

[160] E. Salinas, A. Hernandez, A. Zainos, and R. Romo, "Periodicity and firing and rate as and candidate neural and codes for the and frequency of vibrotactile and stimuli," *Journal of Neuroscience*, vol. 20, no. 14, pp. 5503–5515, 2000.

[161] R. Schaette, C. Turtle, and K. J. Munro, "Reversible induction of phantom auditory sensations through simulated unilateral hearing loss," *PLoS ONE*, vol. 7, no. 6, p. e35238, 06 2012.

[162] L. D. Landau and E. M. Lifshitz, *Statistical physics.* Pergamon Press Ltd., 1980, vol. 5.

A Appendix

A.1 Free energy, the moment-generating function, and their behavior at phase transitions

Here, I take a small excursion to statistical physics to show how the moment-generating function of energy is related to the free energy and why it is helpful in studying phase transitions. If not stated otherwise, information in this section is taken from [90, 162].

The probability of a system to be in energy state E_i is denoted as $p_i \equiv p(E_i)$. Furthermore, the (Helmholtz) free energy of a system is defined as

$$F := \sum_i p_i E_i - TS \ , \tag{A.1}$$

where T is the temperature of the system and S is the entropy, which is defined as (note its resemblance to entropy in information theory, Eq. 2.5 in Sec. 2.3):

$$S := -k_B \sum_i p_i \log p_i \tag{A.2}$$

with Boltzmann constant k_B. The free energy takes into account that a system tends to reduce its energy and increase its entropy. Without outside work, the system can only transition from one state to another if the free energy is not increased during that transition, i.e. if $\Delta F \leq 0$. If the free energy is decreased during a transition, the transition is irreversible without applying outside force. Thus, the system decreases its free energy through irreversible transitions until it reaches a (local) minimum of free energy. Setting $\beta = (k_B T)^{-1}$ and minimizing

$$F = \sum_i p_i E_i + k_B T \sum_i p_i \log p_i \tag{A.3}$$

yields

$$p_i = \frac{e^{-\beta E_i}}{Z(\beta)} \tag{A.4}$$

where $Z(\beta)$ is a normalization term called *partition function*,

$$Z(\beta) = \sum_i p_i e^{-\beta E_i} \ . \tag{A.5}$$

With that, the free energy can thus be expressed as

$$F = -\frac{\mathrm{d}}{\mathrm{d}\beta} \log Z(\beta) \ . \tag{A.6}$$

At the same time, the moment-generating function for the energy is [119]:

$$\begin{aligned} M_E(t) &= \langle e^{tE} \rangle & \text{(A.7)} \\ &= \sum_i p_i e^{tE_i} & \text{(A.8)} \\ &= \frac{1}{Z(\beta)} \sum_i e^{(t-\beta)E_i} \ . & \text{(A.9)} \end{aligned}$$

It is called the moment-generating function since

$$M_E(t) = \sum_n^{\infty} \frac{t^n}{n!} \langle E^n \rangle \tag{A.10}$$

what means that the n-th derivative evaluated at $t = 0$ gives the n-th moment of the energy:

$$M_E^{(n)}(0) = \langle E^n \rangle \ . \tag{A.11}$$

Using Eqs. A.6 and A.9, the relation between free energy and moment-generating function of the energy is:

$$M_E(t) = \sum_n^{\infty} \frac{t^n}{n!} \frac{\mathrm{d}^n(F\beta)}{\mathrm{d}\beta^n} \tag{A.12}$$

Phase transitions occur, when $M_E(t)$ (or, equivalently: the free energy) becomes non-analytic with β (i.e. with temperature). Since the moment-generating function is a sum of exponentials it should be analytic everywhere. Yang and Lee showed that the non-analytical behavior only occurs when the number of particles of the system goes to infinity [121].

Acknowledgements

First of all, I want to thank my supervisor, Julijana Gjorgjieva, who fully supported me during the good and bad times of my PhD and who made a great effort to propel my scientific progress. I could always approach her with all kinds of questions and she also taught me a lot about presenting my work.

I also would like to thank all members of the Gjorgjieva Lab for their continuous feedback on my work and for pleasant times in and outside the office. Special gratitude goes to Jan Kirchner, who proofread my whole thesis.

In addition, I thank people from the Memmesheimer Lab, in particular Raoul-Martin Memmesheimer, Felipe Kalle Kossio, and Sven Goedeke. With them I had many interesting lunch discussions about science, society, and philosophy, and they were also always ready to discuss my scientific problems.

Then, I want to thank my lovely partner, Serena, who has always been supportive and showed a considerable interest in my progress. She also made a great effort to provide me with the perfect writing conditions during the pandemic lockdown.

Finally, I want to express gratitude towards my family for their unconditional support, and in particular, my father and Gitta for proofreading.